THE UNIVERSITY *of* EDINBURGH

Next-generation bioinformatics analysis of bacterial genomes, with a focus on serovar host specificity and pathogenicity in *Salmonella.*

Emily Jane Richardson

This thesis is presented for the degree of
Doctor of Philosophy
University of Edinburgh
2013

Declaration of Originality

I declare that the work described within this thesis and the thesis itself is solely mine unless otherwise stated.

EMILY RICHARDSON

2013

Acknowledgements

I want to thank the people who have supported me throughout; my supervisors, my friends and my family.

The process of completing my PhD and writing my thesis has been a long and difficult one. I particularly want to thank my parents and my good friend Ksenia, they had faith in me even when things seemed so bleak and impossible. I don't think I could have done it without you.

Abstract

Salmonella is one of the most important pathogens of mankind and animals alike, causing several billion pounds worth of damage worldwide each year. We have sequenced, annotated and published 4 genomes of Salmonella of well-defined virulence in farm animals. This provides valuable measures of intraserovar diversity and opportunities to formally link genotypes to phenotypes in target animals. Specifically, we have examined pathway detrition and mutagenesis and linked this to host specificity of the serovars.

With the advent of next generation sequencing there has been a boom in genomic sequence submission, and an onslaught of -omics data has ensued. Integrating these different data types is complex and there is little available to visualise this data in the context of its genome. We present GeneBook, a web-based tool that synchronously integrates disparate datasets, displaying a fully annotated genome, enriched with publicly available data and the user's private experiments. It is accessed through a user-friendly interface that allows scientists to interrogate genomic features across multiple, heterogeneous, experiments.

Publications

- Richardson EJ, Limaye B, Inamdar H, et al (2011) Genome sequences of *Salmonella enterica* serovar Typhimurium, Choleraesuis, Dublin, and Gallinarum strains of well-defined virulence in food-producing animals. *J Bacteriol.* **193(12)**:3162-3 [Genome announcement, IF: 3.726)
- Richardson EJ, Watson M (2012) The automatic annotation of bacterial genomes. *Briefings in Bioinformatics.* doi: 10.1093/bib/bbs007 [Review, IF: 9.283]

Plan/Aims & Objectives

- To ask fundamental questions about host specificity and pathogenicity, focussing on the biology of *Salmonella*
- To annotate novel genomes of recently sequenced *Salmonella* strains and submit to public databank.
- To analyse currently available data for similarities and differences between serovars and elucidate genes associated with host specificity and pathogenicity.
- To provide a flexible bioinformatics platform for the integration of a diverse range of data types available in pathogen biology
- To develop a method of viewing genomes in the context of their disparate data sets which incorporates web services
- To use the developed tool to interrogate both public and private data and demonstrate a method of finding genes of interest that requires little computational expertise.

Contents

Table of Figures

Abbreviations

ABI – Applied Biosystems

AJAX – Asynchronous JavaScript and XML

API – Application Programming Interface

BLAST – Basic Local Alignment Search Tool

BLASTP – Protein-protein BLAST

CDS – Coding Sequence

COG – Clusters of Orthologous Groups

CMS – Content Management System

CSS – Cascading Style Sheets

EBI – European Bioinformatics Institute

ERD – Entity Relationship Diagram

ERIC – Enteropathogen Resource Integration Centre

GEO – Gene Expression Omnibus

GUI – Graphical User Interface

GUS – Genomics Unified Schema

HGT – Horizontal Gene Transfer

HMM – Hidden Markov Model

HTS – High-Throughput Sequencing

HTTP – Hypertext Transport Protocol

KEGG – Kyoto Encyclopedia of Genes and Genomes

MGED – Microarray Gene Expression Data

MIAME – Minimum Information About a Microarray Experiment

NCBI – National Center for Biotechnology Information

ORF – Open Reading Frame

PATRIC – PathoSystems Resource Integration Center

PCR – Polymerase Chain Reaction

PHP – PHP: Hypertext Processor

PSU – Pathogen Sequencing Unit

RDBMS – Relational Database Management System

REST – Representational State Transfer

ROD – Regions of Difference

SOAP – Simple Object Access Protocol

SQL – Structured Query Language

STM – Signature Tagged Mutagenesis

URL – Uniform Resource Locator

W3C – World Wide Web Consortium

WGS – Whole Genome Shotgun

WSDL – Web Services Description Language

XML –eXtensible Markup Language

Chapter One

Introduction

This chapter introduces the core concepts that form the basis of this thesis. Section 1.1 describes the general biology of the Salmonella genus. It goes on to outline the proposed relationship between host specificity and pathogenicity. Section 1.2 explains the current state of sequencing, introducing next generation sequencing (NGS) and some of its challenges. After sequencing, the genomes are annotated, section 1.4 describes generic bacterial genome annotation and some of the limitations of automated annotation. With the increase in bacterial genomics there is an increase in post-genomic data, 1.5 introduces some diverse types of bacterial data and 1.6 explains how this is stored and presented to the public. It presents a critical review of the currently available databases, which store bacterial genomes and their post-genomic data. Finally, section 1.7 looks at the types of web technologies available, it explains the principles of webservices and the various methods available.

The background outlined in this section is designed to form the justification and basis of this thesis. Some parts of this chapter have been published previously in Richardson *et al.* 2011 [1] and Richardson and Watson 2012 [2].

1.1 *Salmonella* Biology

Pathogenic bacteria are of wide economic importance, causing thousands of deaths. In 2009 there were over 10,000 cases of human Salmonellosis in the UK [3], although it is estimated that only a quarter of cases are actually reported [4]. In 2000 there were over approximately 21 million instances of Salmonellosis and more than 200,000 deaths worldwide [5]. Although we see a decrease in human infection, the incidence in livestock such as Cattle and poultry has seen an increase (Figure 1). Controlling the incidence of *Salmonella* is complex because the trends in infection vary between serovars. Reducing the incidence of one type often results in another differentially adapted serovar filling the niche [6]. Initially antibiotic treatment was successful, but indiscriminate use has resulted in the manifestation of multiple drug resistant strains of Salmonellae [7]. These effects have fuelled research into the biology of Salmonellae in an effort to ultimately discover new drug targets and develop methods of disease treatment.

Salmonellae as a genus are promiscuous, infecting a broad range of hosts [8]. The genus is comprised of two species; *Salmonella bongori* and *Salmonella enterica*. The most extensively studied species is *S. enterica* subsp. *enterica* because it is pathogenic to warm blooded animals [9]. The usual mode of infection is through oral ingestion of contaminated food or water [10] resulting in the host contracting salmonellosis. Their ability to cause food poisoning is due to *Salmonella's* capability to colonise the alimentary tracts of livestock, ensuing contaminated carcasses and entry into the human food chain [11].

Within livestock populations some individuals enter a carrier state after clearing an initial infection. They show no clinical signs but still carry and excrete *Salmonella*. into the environment [12], making it difficult to truly eradicate that incidence of *Salmonella* from the population.

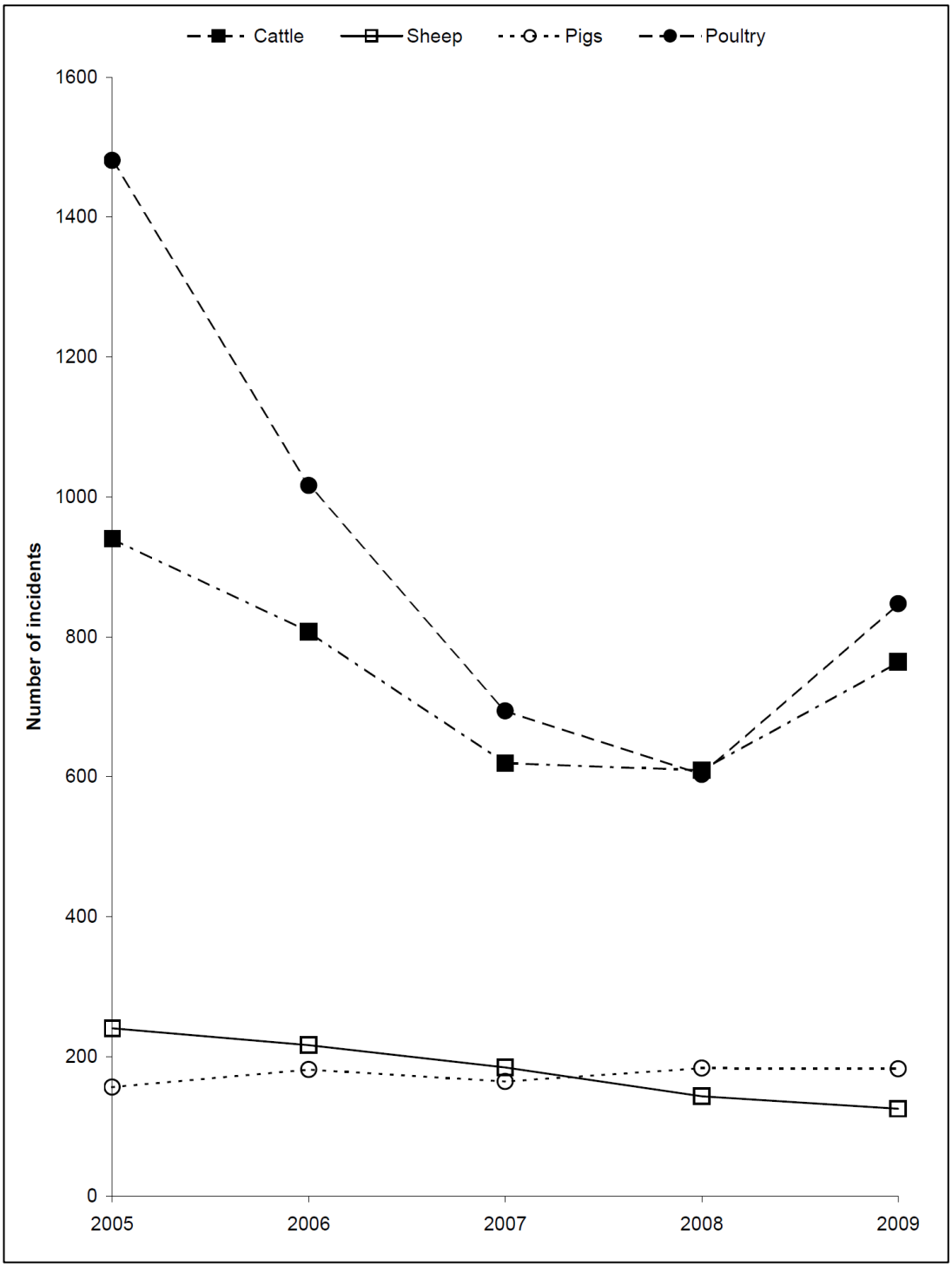


Figure 1 Number of incident reports of Salmonella in livestock according to the VLA's Salmonella in Livestock Production in GB: 2009 Report [13].

Host-pathogen interactions are complex, both party's cells are engaged in the exchange of biochemical signals, resulting in changes to the host cell's physiology to allow pinocytosis of the pathogen cells [14]. A major factor influencing the extremity and symptoms of salmonellosis is the specific serovar which has caused the infection [8]. Systemic infections, such as typhoid, caused by uptake into phagocytic cells can occasionally occur; this type of infection is more severe [15]. The reproductive tract can also be infected by *Salmonella* during the gestation and laying period of mammals and birds respectively. This form of infection is also systemic and can lead to abortion [16]. The other type of infection is caused by *Salmonella* invading gut epithelial cells, leading to gastroenteritis [16].

1.1.1 Characterising *Salmonella*

Within the species *S.enterica* there are over 2,400 serovars. Each serovar is distinguishable by its serology, that is the variation of surface exposed antigens (O,H and Vi) [17]. Salmonella diversity doesn't end at the serovar level [7, 18]. For example, Typhi strains vary in their drug resistance, number of plasmids, prophages and island structures [7]. The use of serology as a method of identification can only differentiate strains with a limited level of success. In 2004 Porwollik *et al.* used microarray analysis to reveal strains within the same serovar which differ by hundreds of absent/present genes [9, 19]. They proposed to group strains based on their gene profile, namely 'genovars'. The fact that multiple genovars can appear in one serovar suggests that genes coding antigens and flagella are passed via horizontal gene transfer (HGT) [9].

There is an essential core gene set which all *Salmonella* share. Jacobsen *et al.* tried to define this along with the Salmonella pan genome by comparing 45 publically available Salmonella genomes (21 serotypes) [20]. They estimated that the core genome consisted of 2800 genes and that the pan genome had at least 10,000 genes [20].

Each serovar has a niche specific sub-gene pool, which can vary dramatically within serovars; older serovars (e.g. Typhimurium) generally being less homogenous than

the more recently diverged (e.g. Typhi) [21]. The homogeneity of serovars like Typhi can also be attributed to their restricted host range. A recent comparison of 19 strains of Typhi revealed scarce adaptive selection except in the gyrA gene which is associated with fluoroquinolone resistance [22].

1.1.2 Host Specificity

The range of hosts that *S.enterica* subsp. *enterica* can infect is broad. Each serovar varies in its host range, specificity and pathogenicity. There are serovars which can infect many hosts, these are host ubiquitous (e.g. serovar Typhimurium). Some serovars can only infect a narrow selection of hosts, these are host-restricted (e.g. serovars Dublin and Choleraesius). Finally some serovars can only infect one host, these are host specific (e.g. serovars Typhi and Paratyphi A). The use of serovars to distinguish host specificity is not failsafe, there is evidence that some strains of the ubiquitous serovar Typhimurium have a very narrow host range [23].

In general terms the specificity of a strain is linked to its pathogenicity; host-generalists tend to cause acute but self-limiting gastroenteritis with little systemic involvement in healthy outbred adults compared to host restricted strains that cause systemic disease [24]. Lawrence *et al.* outline the evolution from a broad host range to host-dependency (Figure 2, [25]). Unfortunately, livestock populations are not purely comprised of the model healthy outbred adults. For example, young animals are especially vulnerable to systemic infection, regardless of the accepted host specificity of the strain [12]. Also, latent carriers can show clinical signs again when under stress [12].

Comparative genomics has demonstrated that host-specialists evolved from host-generalists. A comparison between the chicken restricted serovar Gallinarum and serovar Enteritidis has determined that Gallinarum is a recent descendent of Enteritidis. It is primarily through genome degradation that Gallinarum has become host restricted [26]. Comparisons between two different human specific serovars, namely serovar Typhi and serovar Paratyphi A revealed they are similar genetically [20, 24]. McClelland *et al.* in their 2004 study found that they not only shared pseudogenes they also showed disruption of different genes that reside in the same

pathway. Host specifist gene loss is often linked to motility or fimbriae formation [27], it is suggested that this degradation may augment the evasion of TLR-5-induced pro-inflammatory responses of the host [26].

Genome degradation is linked to pseudogene formation. Pseudogenes are formed by mutations, leading to either a frameshift or a stop codon for example. High levels of pseudogenes imply gene loss, possibly removing genes which are redundant since adhering to a host-specific niche [24]. Identifying pathogenesis-associated genes that are common between two highly degraded genomes indicates their importance in enteric disease and highlights potential target genes [24]. Loss of gene function can be beneficial on an economic level, where oral vaccines are developed from strains which lose their pathogenicity through the down-regulation of virulence genes [28]. It is worth noting that on occasion gene acquisition also plays a part in host-specific adaptation. The viaB locus is present in *S.* Typhi but not in *S.* Typhimurium it is believed to prevent detection of pathogen lipopolysaccharides by the TLR4-mediated host response [29].

The other driving force behind niche adaptation is HGT. This produces quantum leaps in evolution allowing exploitation of new niches and rapid adaptation to environmental pressures (such as host antibiotic use) [25]. It is estimated that over a quarter of the Typhimurium genome is from HGT [30].

Genomic islands are products of HGT and many are associated with pathogenicity, namely *Salmonella* pathogenicity islands (SPIs). These can be identified by their low G+C content compared to the rest of the genome [31]. Salmonellae are able to manage HGT in the genome by selectively repressing transcription in areas with a low G+C content [32].

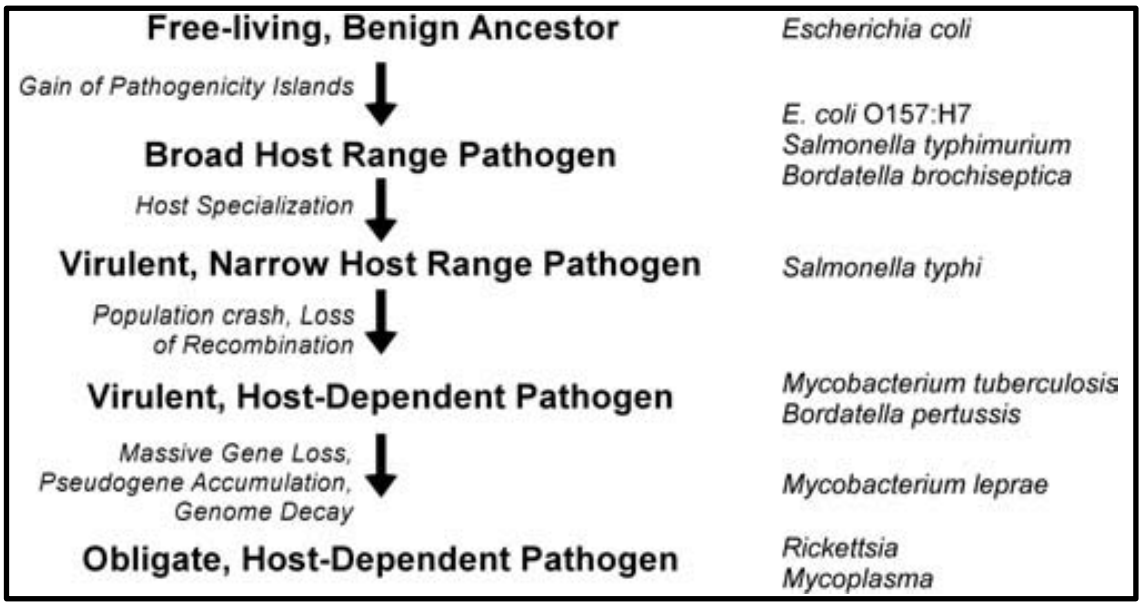


Figure 2 shows the effects of horizontal gene transfer and gene loss on increased host dependency in pathogenic bacteria [25].

1.2 Genome Sequencing

With the arrival of ultra-high-throughput sequencing, we are in the midst of a revolution in comparative genomics. The success of this is partially dependant on suitable tools and analysis techniques [33]. Comparative genomics is a major aspect of pathogen studies. It is often used as a basis for lab based experiments, identifying candidate genes and giving a biological insight into niche adaptations. The number of fully sequenced genomes available in GenBank, the National Center for Biotechnology Information's (NCBI) sequence repository, is soaring (Figure 3 [34]).

Inevitably, with the proliferation of sequenced genomes, the post genomic revolution has ensued. This increase in information has resulted in the release of hundreds of new databases and datasets every year. For example, Nucleic Acids Research publishes a database edition every year, which records many of the biological databases that have been released that year. The 2009 edition is recorded as having over 1000 databases available [35]. Furthermore, Oxford University Press have recently announced a new journal entitled 'Database: The Journal of Biological Databases and Curation' [36]. The journal is dedicated solely to biological databases and "aims to help strengthen the bridge between database developers and users" [36]. The increase in database publications demonstrates the importance of data management and integration in modern biological research.

Not only is there much repetition across these databases they also reside in many different locations. This makes it difficult for scientists to find all the information relevant to their research interests. When they do find it the fact that it is in multiple locations makes extracting anything useful an expansive task consisting of multiple open tabs in the browser and running various tools, some locally and some via the web.

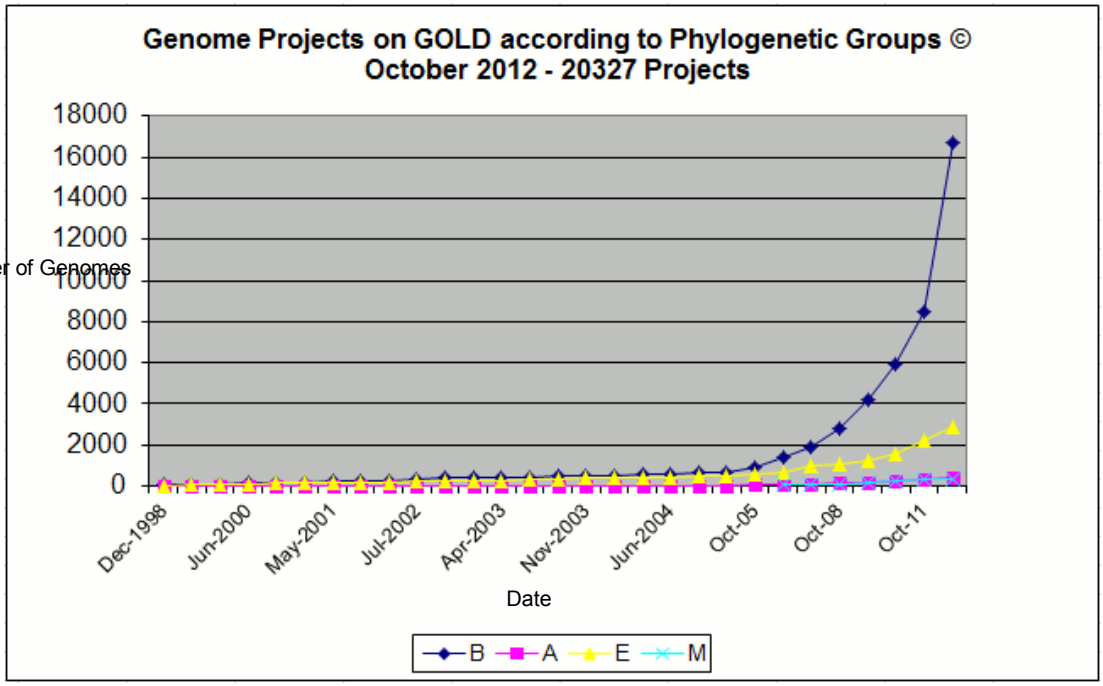


Figure 3 showing the number of sequenced genomes submitted to GenBank between 1998 and February 2012. The legend describes different kinds of genomes; Bacterial (blue, B), Archael (pink, A), Eukaryotic (yellow, E) and Metagenomic (cyan, M). [34]

1.2.1 Advancements in Sequencing Technology

In 1977 Frederick Sanger and his colleagues announced a new method for DNA sequencing, later called 'Sanger Sequencing' [37]. Recently we have seen a new generation of technologies that have strived to become faster and cheaper. In the past decade genome sequencing has become a lucrative business, and several new companies have emerged with their own sequencing technologies, collectively referred to as next generation sequencing (NGS). There are papers and reviews available [38-42], but these rapidly become out of date as the pace of change in DNA sequencing can be faster than the peer review process. Information for this section is taken from the above reviews, from personal communications with sequencing experts within ARK-Genomics, and from my own knowledge gained from reading blogs, forums, and papers and from attending conferences. Below is a review of the current platforms available, which have been summarised in Table 1.

The revolution in DNA sequencing began in 2005 with 454 Life Sciences, bought by Roche in 2007. The 454 technology is a sequencing-by-synthesis technology whereby molecules of fragmented DNA are bound to beads which then undergo emulsion PCR. Each bead therefore represents multiple copies of a single strand of DNA. These beads are arranged in wells on plates, and single nucleotides are washed across the plate in sequence. As each nucleotide is incorporated, a fluorescent label, coloured according to base, binds to it preventing further amplification. A laser is projected across the wells and the fluorescence is captured by a camera. A chemical wash then releases the dye allowing the incorporation of the next base. Images are built up sequentially and when analysed, these images can be translated into the underlying sequence in each bead. The average read lengths on the Roche 454 FLX machine are approximately 400-500bp. A recent advance is the 454 FLX+ which produces reads of approximately 700bp. The number of reads obtained for both FLX and FLX+ is typically 1 million.

Competing technologies came from Solexa, later bought by Illumina, and ABI's SOLiD system. Both of these systems are also sequencing-by-synthesis, single strands of DNA are used as a template and the technology measures incorporation of bases into the second strand. Both Solexa and SOLiD began as ultra-short read

technologies, producing reads around 35bp. Solexa's technology is based around bridge amplification. Fragmented single-stranded DNA has adapters ligated, and is attached to a slide. Due to adapters attached to the slide, each single strand is able to form a loop between two adapters, and a second strand is built. The two strands then separate, the process is repeated, and in this way the original molecule is amplified many times. This process is called bridge amplification, and results in clusters of single stranded DNA. After cluster generation has occurred, the four bases are again washed across the slide, and the fluorescence is measured by a camera after each base is incorporated. The images produced by the camera can then be analysed to deduce the underlying sequence. The Solexa technology underwent huge advance after being bought by Illumina, with read lengths now up to 250bp on the MiSeq, and reads-per-lane up to 180 million on the HiSeq.

The ABI SOLiD system measures DNA sequence in "colour space". Similar to the 454 technology, fragmented single stranded DNA is bound to beads and undergoes emulsion PCR such that each bead represents only one DNA molecule in multiple copies. Bases are then incorporated as dinucleotides with each dinucleotide being represented as one of four colours. The reaction then moves along one base of the template and is repeated. Thus each base in the template is measured twice, by two different dinucleotides/colours. Reads are provided not as bases but as a sequence of colours, which can be translated into "base space". Reads are generally quite short, with a maximum of 75bp, and approximately 50 million reads per lane.

Each of the three major companies also provide benchtop sequencers that allow rapid analysis of DNA samples. The Roche 454 Junior employs the same technology as the 454 FLX system, so reads are on average 400bp and the junior produces about 80,000 reads. The Illumina MiSeq employs the same technology as the Illumina HiSeq, but due to a smaller slide size and improved fluidics, longer read lengths are possible, currently up to 250bp. The MiSeq can produce approximately 13 million reads.

In 2010, ABI/LifeTech bought Ion Torrent, a small sequencing company created by Jonathon Rothberg, the inventor of 454. Ion Torrent's technology is similar to that of 454, except instead of measuring the release of light during the incorporation of bases, a semiconductor measures the release of Hydrogen ions. The Ion Torrent system is very quick in terms of sequencing, and can produce 200-400bp reads and up to 3 million reads per run.

Table 1 Summary of current next generation sequencing platforms. Where N is a general platform, B refers to the smaller benchtop machines and S are single molecule sequencing (SMS) platforms.

Machine	Type	Read lengths	Chambers/Lanes	Reads per chamber/lane
HiSeq 2500 traditional)	N	35, 50, 75, 100	16	150-180,000,000
HiSeq 2500 (rapid run)	N	50, 150	2	150-180,000,000
Roche 454	N	400 (mode)	1	1,000,000
ABI SOLiD	N	35, 60, 75	12	50,000,000
Ion Torrent	B	Up to 400	1	2-3,000,000
MiSeq	B	50, 150	1	13,000,000
454 Junior	B	400 (mode)	1	80,000
PacBio	S	2.5-5kb (mean)	1	35-80,000
Helicos	S	25-70	100	12-20,000,000

All of the above technologies are limited in that they rely on PCR, or amplification, of the DNA template. Third generation sequencing is already available, still in its infancy, 'Single-Molecule Sequencing' (SMS), Gupta summarises SMS in her 2008 review [43]. There are two so-called single-molecule sequencing technologies that do not rely on amplification and sequence DNA and RNA directly.

With the Helicos heliscope sequencer, single-stranded fragmented DNA is created with polyA tails and attached to a flow cell surface. The incorporation of fluorescently labelled nucleotides is then measured sequentially by camera and records the incorporation of the second strand into each single molecule. Read lengths are limited to 55 base pairs and the system is capable of 1 billion reads per run. Unfortunately, the system suffers from quite high error rates, 1-3% indel rates, which coupled with the short reads, can cause problems for mapping to reference genomes. The system is the only technology to date to demonstrate direct sequencing of RNA (without conversion to cDNA).

Finally, Pacific Biosciences also produce a single-molecule sequencing technology. Here, a version of the DNA polymerase enzyme is placed at the bottom of a well, and single strands of DNA fed into it along with fluorescently labelled single nucleotides. The system measures the fluorescence as the DNA polymerase incorporates bases into the template. Reads can average 2-3kb, with ultra-long reads possible, notably, a 17kb read has been reported. There are 80,000 wells though not all produce a useable sequence. The error rate is quite high, 15% has been reported, and many of these are indels. The error rate, coupled with the low throughput, has prevented the wider adoption of this technology. Accuracy can be increased at the expense of read length by circularising the template and measuring each base multiple times.

With these advancements whole genome sequencing is no longer restricted to sequencing centres; any research group with a small amount of funding can pay to sequence their genome of interest. The fact that this is process is becoming quicker and cheaper at an astounding rate means that there will be a more genomes which

need annotation. Further to this 'desktop sequencers' have been developed by Illumina, Solexa and Ion Torrent (MiSeq, 454 GS Junior and ION Torrent IGM respectively). These are designed to be held in any lab, even hospitals for quick and relatively inexpensive sequencing. This is especially exciting for bacterial genomics as these machines can sequence an entire genome in 24 hours or less [42]

On the horizon, and predicted to be a big contender, is nanopore sequencing, which uses nanoscale technology to produce larger reads in real time. That is, the process of sequencing is physical rather than chemical, feeding DNA strands through nanopores and reading the sequence molecule by molecule [44]. This technology is predicted to be much faster and produce considerably longer reads. Companies like ON (who are closest to market), IBM and Visigen have previously made announcements predicting that they will be offering single molecule sequencing services in 2010, at this time (2012) no nanopore technologies have been made available to the public [44, 45].

1.2.2 NGS and bacterial genome sequencing

The advances in NGS technologies have changed the state of bacterial sequencing. The luxury of having a relatively small genome (compared to Eukaryotes) means that it is actually possible to sequence a complete bacterial genome for less than £100 in under 24 hours. As a result of these benefits we see a movement towards whole genome scale analyses. The fact that analyses are on the genomic scale allows for more holistic comparisons between strains. For example, it is possible to identify mobile genetic elements and look for all single nucleotide polymorphisms SNPs rather than regions restricted to a microarray. Sometimes the differences between strains of bacteria can be due to SNPs. These single base changes can actually result in different amino acids and sometimes different phenotypes [1]. Mostly, SNPs are synonymous, that is the amino acid sequence remains intact. The identification of non-synonymous SNPs can be indicative of niche adaptation. In terms of bacterial evolution all SNPs (synonymous and non-synonymous) can be used as a means of looking at the phylogeny of strains [22].

SNP detection can be useful for distinguishing between strains, this can be for epidemiological traceback (to the origin of an outbreak for example) and for insights into ecological niches (such as host adaptation). Traditionally, for subtyping isolates pulsed field gel electrophoresis is used. However for highly related clones the PFGE pattern is often shared between strains making outbreak origin difficult to elucidate. Whole genome sequencing can be used for single nucleotide polymorphism prediction, which can be used as a basis subtyping of strain [46],[47] . In their 2012 study Allard *et al.* examined the SNPs for many serovar Montevideo isolates [48]. They showed that the isolates are well conserved using PFGE regardless of their geographical/temporal differences. However, when using WGS comparisons between isolates from an outbreak other isolates, not associated with the outbreak, there were only a few SNPs difference within the outbreak isolates and up to thousands of SNPs difference between the other isolates; demonstrating that the NGS approach shows strain differences at a much higher resolution.

As many bacteria are pathogenic, one of the uses of NGS in terms of bacterial sequencing is outbreak analysis. A vision for desktop sequencing in particular is the use in diagnostics; a pathogen can be isolated from a patient and sequenced within the hospital, giving the medical staff a clear idea of what they are dealing with and a means to decide on how to administer care based on the specific infection. During the 2011 *E. coli* outbreak there was a worldwide collaborative effort to solve why strain O104:H4 was so virulent. Different laboratories were sequencing outbreak isolates and then making the output publically available [49, 50]. Collectively they identified that the strain carries a shiga toxin gene, known for its haemorrhagic properties. This was only possible due to the low cost and quick turnover of NGS technologies.

Comparative genomics is enhanced by NGS, the whole genomes of serovars can be used to look at the divergence of strains. Looking at the genomic scale allows for comparisons between recombination, loss of gene function and distribution of SNPs. This type of research has elucidated the phylogeny of strains with low genetic variation, it has been linked to the divergence of host adaptive traits with the

divergence of the host into domestication [22] [51] and further to this NGS has allowed Salmonella to be divided into two distinct clades based on SNPs and HGT and demonstrated the genetic differences isolates of the same strain that show different egg colonisation abilities [51].

Just the fact that NGS is so accessible in terms of cost and time and the fact that many sequenced strains are made publically available means that we can look at pan genomes [20].Studying these allows us to get a better understanding of what defines that species/strain/serovar which ultimately can allow us to identify what results in what phenotype. The Salmonella pan genome has been compared to the E. Coli pan genome, identifying potential species specific gene profiles [52]. This type of analysis can be used as a basis for finding the origin of genes, that is whether they were a result of HGT or from a common ancestor [52].

1.2.3 NGS genome assembly

After an isolate has been sequenced some kind of genome construction usually occurs. Broadly speaking this can be performed in one of two ways, either mapping the reads to an existing reference genome or using an *ab* initio method to assemble the reads without a reference, namely reference alignment and de novo assembly. There are many papers which have performed comparisons of these but *Magoc et al.* use real data for a very extensive comparison [53].

1.2.3.1 Aligning reads to a reference genome

In terms of bacterial sequencing there are many novel and resequencing efforts. Being able to map the reads to the original sequence allows for simple detection of errors and resolving repeat regions.

The premise of read mapping is to use a reference as a back bone and align the reads iteratively to this resulting in a pileup of reads across the genome. From this a consensus sequence can be made which calls the most likely base for each nucleotide location in the genome. The base is decided based the nucleotide which occurs most frequently at that location and a quality score is assigned depending on the frequency of that base at that location.

As the reads are actually mapped against the reference the coverage for each base location in the reference is available. This can help with distinguishing SNPs (bases that differ from the reference) and sequencing errors. High coverage means that a location where there is a sequencing error will also be saturated by reads that have the correct base meaning. For example if the location has 50x coverage and 48 of the reads predict one base it is a fair assumption that the two bases with a converse base are either sequencing errors or a mismapped read.

1.2.3.2 De novo assembly

In cases where there is no reference genome available or when regions of difference (such as mobile genetic elements and plasmids) are being identified mapping reads to a reference is not a suitable method. In these scenarios de novo assembly is required, most assemblers rely on graph based methods such as de bruijn graph. These build up the reads (or fragments of the reads) into bigger structures called contigs. If a reference genome is available the contigs can be mapped to this (using a tool such as abacas) in order to build a scaffold of the whole genome and look for regions of difference.

1.3 NGS Challenges

The impact that NGS technologies have had on research is titanic, but that is not to say that next generation sequencing does not have its own limitations and challenges.

Current NGS datasets are notoriously large with a single lane of HiSeq providing 6000x coverage of a 5Mb bacterial genome and the average bacterial alignment file being many gigabytes in size. This causes problems for software that tries to load the entire dataset into RAM, thus, many require a 64 bit operating system with larger amounts of RAM. Tools such as Samtools have been developed for dealing with alignment data, enabling users to convert their data to binary format and use fast random access of alignment data [54]. Tools such as IGV and Tablet are available for visualising this type of data [55, 56], but the sheer amount of information in one file makes visualisation difficult whilst displaying regions of the sequence, they are also liable to crash if the region viewed is too large or has very high coverage. The promise of third generation sequencing will come to fruition, and the idiosyncrasies

associated with current 'NGS' processing and analysis will no longer apply, the new technology will have its own limitations. Although, the advent of single molecule is imminent we are still a long way from using this technology in day-to-day analysis, new software will need to be written to handle and analyse the new data types, new databases will need to be made for the storage. With that in mind the bottleneck will be the computational post sequencing aspect of the process.

1.4 Genome Annotation

Prokaryotic genomics has seen an explosion in the number of genome projects, driven by the advent of next generation sequencing (NGS), resulting in a huge reduction in the time and money investment per project [57] Microbial genome annotation often consists of running an automatic annotation pipeline followed by manual curation of the results [58]. Most annotation pipelines use homology methods to transfer information from a closely related reference genome to the new sequence. Automatic pipelines can lead to the introduction and propagation of poor annotation and errors, and it is the purpose of the manual curation step to identify and remove these. However, as it is now possible to sequence multiple microbial genomes in a single day at low cost using a single sequencing machine [3], it is no longer feasible to manually curate the annotation of all sequenced genomes. Fully-automatic annotation pipelines, while essential to the modern microbial genomicist, may introduce and propagate inconsistent and incorrect gene annotations.

Transferring annotation purely based on the closest annotated relative does have its limitations. When we consider the reason the new strain has been sequenced, often it will be to identify how this strains differ genetically to its close relatives. This is paradoxical because we are trying to find the differences between these strains but using a similarity based method to annotate it. Potential areas of interest may not be annotated because they are not in the reference genome.

1.4.1 Generic annotation process

Upon sequencing a genome the next step is gene annotation, which forms the basis for many comparative studies. Gene prediction tools are used to ascertain the location and probable function of genes [59]. Two important tools are GLIMMER,

and GeneScan [59, 60]. These techniques apply *ab initio* methodologies; that is they predict genes based on patterns in the sequence rather than homology with other sequences. Using homology, when available, can enhance gene prediction as highly similar sequences are likely to code for the same gene. Orpheus uses both *ab initio* and homologs with apparently high levels of prediction [61].

Once the gene locations have been determined, they are annotated by homology searches to databases such as Pfam [62], NCBI gene [63], and the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [64]. Orthologous gene relationships between species are also calculated and stored in databases such as KEGG and Clusters of Orthologous Groups (COGs) [65].

A range of automatic bacterial annotation pipelines have been published, including web-based systems such as RAST [66], BASys [67], WeGAS [68] and MaGe/Microscope [69]; and systems to be locally installed, such as AGeS [70], DIYA [71] and PIPA [72]. There is also MICheck [73] which checks annotated sequences for syntactic errors. All of these systems carry out the basic process outlined above, with various additions to check for errors or add additional information. It is worth noting that in order to submit to a genome repository that the annotation needs to be in a compatible format (e.g. .tab or .asn). Some pipelines do not output in this manner as they are designed to either hold the annotation online or for in-house analysis [67], [68]. Further processing may therefore be necessary before submission to a public database.

In many cases there is a closely related strain/serovar available which has already been sequenced and annotated. Most annotation pipelines employ gene prediction software, the most common of which is GLIMMER [60]. This uses a reference set of sequences to train a model and then utilises that model to predict coding regions in the genome of interest. Many other *ab initio* gene prediction algorithms exist and these are reviewed by Do and Choi [74]. Alternatively, gene finding can be performed by extrinsic methods, identifying open reading frames directly from comparisons to protein databases [75, 76].

Once coding regions have been identified, they are aligned either to a reference genome annotation or the entirety of UniProt [77] using fast sequence alignment tools (e.g. FASTA [78] or BLAST [79]), the top hits are accepted as homologues and the annotation is transferred across for genes displaying high similarity. Other features such as tRNAs and rRNAs may then added using other prediction software [80].

Here we describe a very general process used for bacterial genome annotation (Figure 4). A more thorough review can be found in Stothard and Wishart [58].

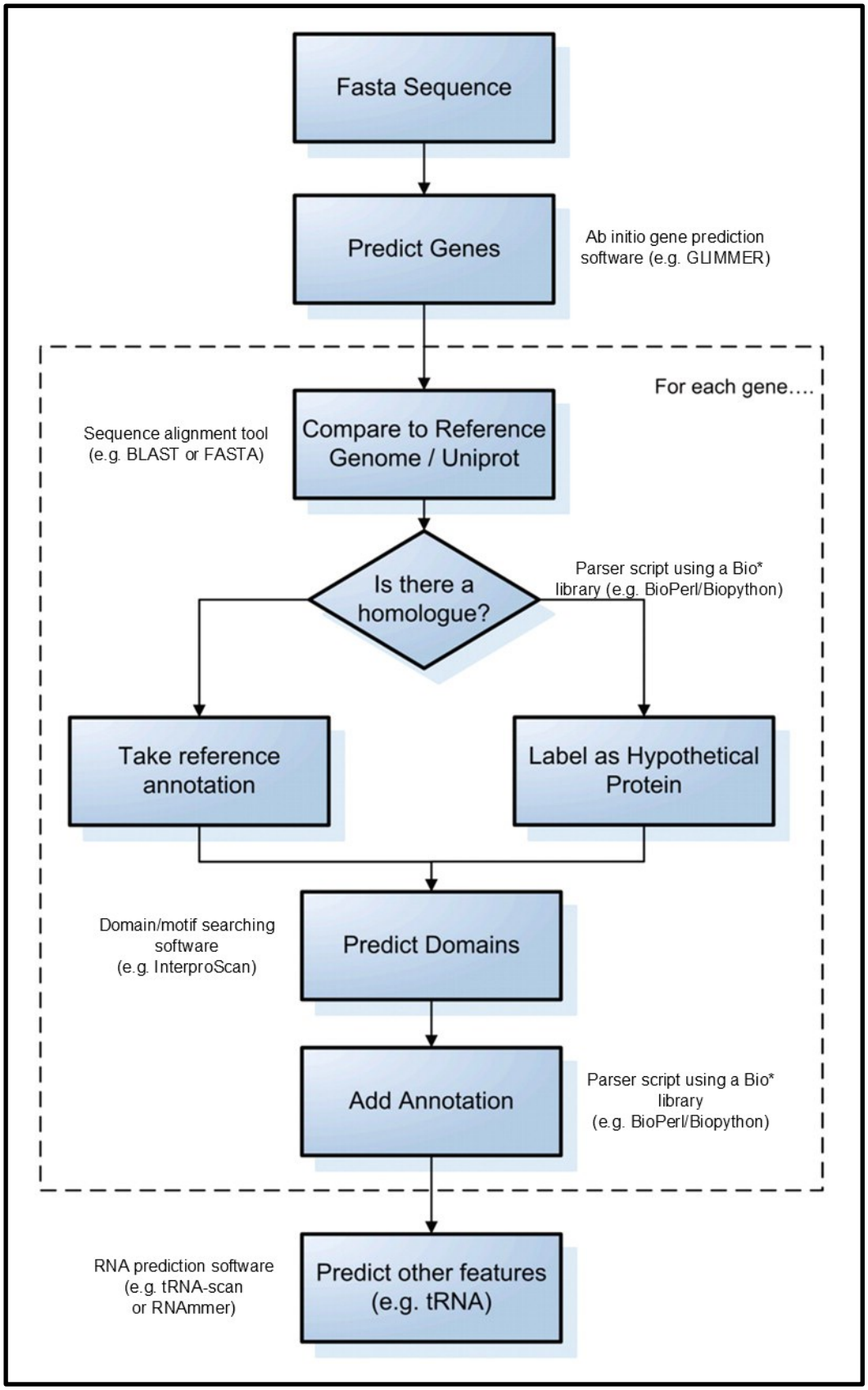


Figure 4 A generic process for automated bacterial genome annotation, the dashed box shows iteration for each gene in the genome, section 2.2.3 describes a working example of the annotation process.

1.4.2 Limitations of the annotation process

In an ideal world this would be the end of the annotation process. The fact that homology is the basis for these pipelines means that many genomes currently available may have been annotated using old, out of date genomes as a reference which in turn have been annotated based on even older more out of date genomes.

The misannotations and errors may perpetuate throughout each new genome, ultimately propagating into secondary databases such as UniProt [77] and KEGG [81].

The public sequence databases have recognized the need for controlling this replication of errors and provide validation software for checking the standard of one's annotation prior to submission [82, 83]. This section looks at common errors that are the product of automated annotation and tries to address methods of overcoming these.

Many bacterial genera now have multiple species and strains with complete genomes, representing a fantastic resource for comparative genomics. However, each genome is annotated separately, by a range of different groups using different protocols, and this introduces inconsistencies. Meaning that reference genome choice can actually result in different numbers of genes predicted/annotated [20]. One particular problem is that of split/fused genes and domains; Kummerfield and Teichman [84] found that, of 7116 distinct domain architectures examined across 131 archaeal, bacterial and eukaryotic genomes, 47% showed evidence of gene fusion/fission events. An example of this is the eutM/eutN locus in *Salmonella*. Figure 5 shows six different models that have been used to annotate this region in the 17 RefSeq records for *Salmonella* in 2012. In *Salmonella typhi* CT18 (NC_003198) and *Salmonella typhi* Ty2 (NC_004631) there is a single ORF of 690bp annotated as eutN (Figure 5A). The protein sequence maps to two domains in PFAM, a BMC domain (PF00936) and a eutN_CcmL domain (PF03319). In all other *Salmonella* genomes in RefSeq, stop codons within this region split the gene, and the domains, in two. In one genome (NC_012125) the region has been annotated as a single long

pseudogene of 690bp (Figure 5B); a further four genomes annotate two intact gene/CDS features, eutM and eutN, each ~300bp in length (Figure 5C). A further three genomes are annotated with one pseudogene, a 291bp ORF equivalent to the eutM gene in Figure 5C, and one intact gene, a 288bp ORF labelled as eutN (Figure 5D). A further two genomes annotate two ORFs, 291bp and 300bp in length respectively, both annotated as pseudogenes (Figure 5E), equivalent to the eutM and eutN genes in Figure 5C. Finally, one genome (NC_006511) includes two intact genes, but has reversed the order of eutM and eutN (Figure 5F).

The various ways in which the eutN and eutM genes have been annotated represents a problem for further genome annotation. We cannot know, simply from the genome sequences alone, whether this locus represents a single long gene that has been split in two, or two shorter genes that have become fused. All six models represent different interpretations of a locus that is highly conserved at the nucleotide level across *Salmonella* species, and any novel genome that is compared to just one of those models will have annotation heavily influenced by that model. For example, if a novel genome is compared only to genomes represented by Figure 5B (two short ORFs annotated as a single long pseudogene) the interpretation will be very different than if the genome were compared to Figure 5C (two short ORFs annotated as two separate intact genes).

Predicting domains directly, rather than genes, using tools such as PfamAlyzer [85], may help in regions with split genes. In the case of eutM/eutN in Salmo*nella*, a domain search would identify two intact domains in all cases; however, the question of whether or not those domains come from the same or separate genes would remain unresolved. We are left with two different versions of the eutN gene from *Salmonella* in the public databases, one of 690bp containing two domains, and one of ~290bp with one domain.

The only way to annotate this region correctly *in silico* would be to compare any new genome to each of the six different models. It is difficult to imagine a set of rules that could be given to an automatic annotation pipeline to interpret correctly the evolution

of this region and apply that interpretation to a newly sequenced genome. To truly get the full story we would need to look at experimental data (such as RNA-Seq data) to see what the patterns of expression are. Sections 4.3.1.4 and 4.3.1.5 show examples of how integrating different data types can help to decipher this kind of annotation problem.

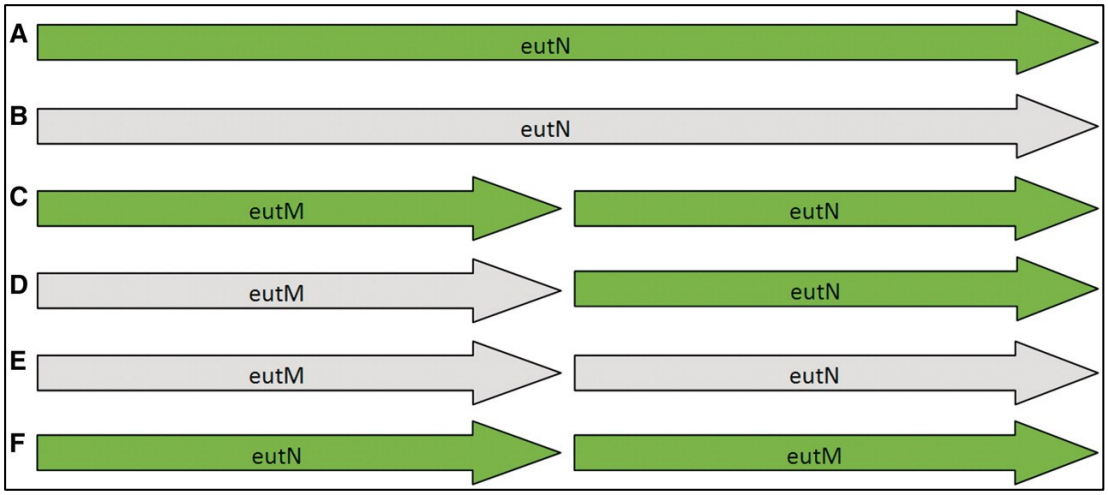


Figure 5 The six different models present across 17 RefSeq entries for Salmonella species at the eutM/eutN locus. Green indicates normal gene/CDS features, grey indicates gene features annotated as pseudogenes. (A) A single intact gene of 690bp; (B) a single pseudogene of 690 bp; (C) two short intact genes ~300bp in length; (D) one pseudogene and one intact gene, each ~300bp in length; (E) two pseudogenes, each 300bp in length; and (F) two intact genes with the order reversed.

In the eutN/eutM example, we see a case where genes of vastly differing lengths have been given the same gene name in different genomes; in contrast to this, it is also possible for orthologous genes to be assigned different gene names. Figure 6 shows a syntenic block of genes annotated in *Escherichia coli K12 MG1655* (NC_000913) and *E. coli O157:H7 Sakai* (NC_002695). These two regions are more than 97% identical at the nucleotide level; however, the annotation differs considerably. While *E. coli K12 MG1655* contains features with gene names araA, araB and araC, the equivalent features in *E. coli 0157:H7 Sakai* do not have those gene names and have been assigned uninformative locus tags. Further information is available for the features with only locus tags, including their involvement in arabinose metabolism, however, the gene names remain absent. At the far right of the gene block, two orthologous features exist, both with gene names, however, this time the problem is that they are different: thiB in *K12 MG1655* and tbpA in *0157:H7 Sakai*. A simple search of the NCBI gene database (search term 'thiB AND Escherichia coli [Organism]' versus search term 'tbpA AND Escherichia coli [Organism]') reveals that both features code for a thiamin(e) transporter subunit, but the gene is given the gene name tbpA in over 30 *E. coli* species, whereas it is given the name thiB in only one. Luckily, the thiB feature in *K12 MG1655* lists tbpA as a 'synonym'. Finally, in the centre of the image, *K12 MG1655* contains a feature with the gene name yabI, whereas its ortholog in *0157:H7 Sakai* only has a locus tag. This is an example of a y-gene, which we discuss in greater detail in the 'Hypothetical proteins' section, 2.4.2.3.

The major issue here is that not only do different genomes annotate orthologous genes differently, and provide inconsistent information; they also contain differing amounts of information. This means that, when annotating a new genome, it is essential to choose a reference genome that contains the most accurate and up-to-date information, and that it is also preferable to compare any new genome to multiple references such that inconsistent annotations can be identified and resolved.

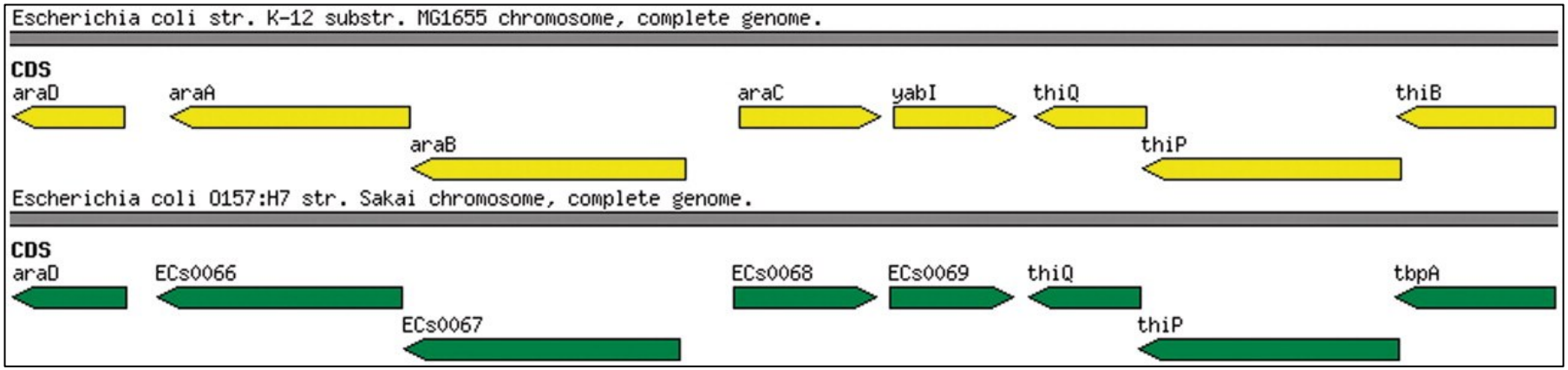


Figure 6 A syntenic block of genes showing inconsistent gene name annotations in *E. Coli* K12 MG1655 and *E. coli* 0157:H7 Sakai.

The definition of orthologous and paralogous genes is of great importance when annotating novel genomes. Whereas 'homology' refers to genes that simply share a common origin, 'orthology' refers to genes that arise by speciation and 'paralogy' refers to genes that arise by duplication. Figure 7 shows some of the processes that can lead to, and define, orthologs and paralogs. Beginning with a single ancestral, a gene duplication event occurs to create two paralogous genes. After a speciation event, there are two different organisms that both contain the paralogous genes from the gene duplication event. Gene 1a in Organism 1 has three homologs after the speciation event. Gene 1a in Organism 1 and Gene 1a in Organism 2 are orthologs as they have only been separated by the speciation event. Gene 1a in Organism 1 and Gene 1b in Organism 1 are in-paralogs, as they have only been separated by the gene duplication event. Finally, Gene 1a in Organism 1 and Gene 1b in Organism 2 are out-paralogs, as they have been separated by the gene duplication and the speciation event.

These processes are not only crucial in defining evolutionary relationships, but also functional relationships, as orthologs tend to retain similar functions, whereas paralogs tend to diverge over time to perform different functions (reviewed in [86]). Therefore, when transferring functional annotation from a sequenced genome to a novel genome, it is essential that orthologs are accurately defined. There are several computational approaches which can be used to accurately define orthologs (reviewed in [87]). Phylogenetic tree-based approaches attempt to reconstruct the evolutionary relationship between gene sequences and thus define orthologs and paralogs; however, it may be impractical to construct a phylogenetic tree for every gene in a newly sequence genome. An alternative is the "bidirectional" or "reciprocal" best-hit approach [88], usually determined by comparing the top-ranking matches found by a search algorithm such as BLAST or FASTA [78, 79]. Gene Synteny, the conservation of local gene order, can also help distinguish orthologs from paralogs in closely related genomes. However, it is important to note that a number of processes can lead to the breakdown of absolute gene synteny, resulting in genuine orthologs having a different gene order. These processes include gene duplication or fusion events, local rearrangements (insertions/deletions) and

translocations. It is important that we model these processes to allow the correct identification of orthologs in complex cases, and the MaGe [69] system attempts to do this. Finally, it has been observed that orthologs exhibit a greater level of protein domain architecture conservation than paralogs [89]. In practice, it may be essential to use a combination of approaches, and several software applications exist [87].

The process of running an annotation pipeline is straightforward. However, there will be errors in the annotation. This section has outlined some of the shortcomings of automatic annotation. Some of the annotation discrepancies such as ortholog/paralog, pseudogenes and gene fusion events can be resolved by integrating the annotation with experimental data such as mutagenesis and microarray.

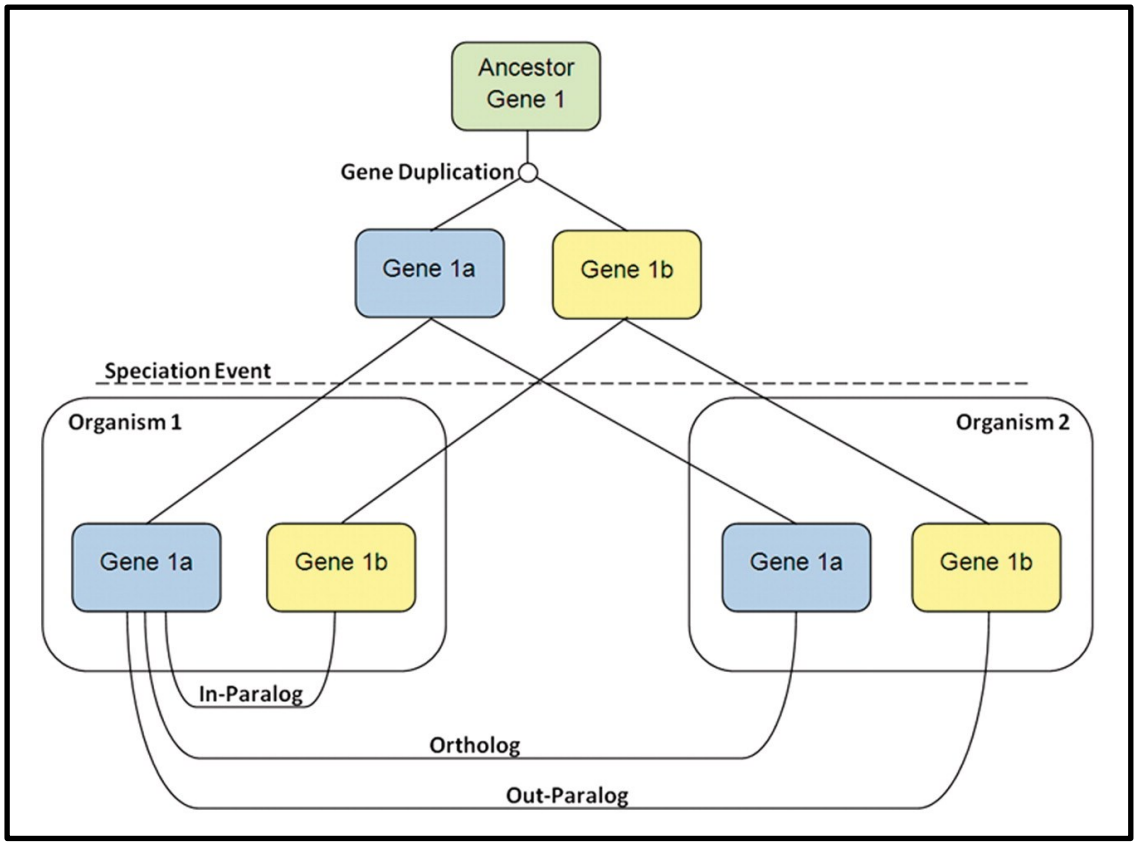


Figure 7 The processes that can lead to, and define, orthologs and paralogs. Gene duplication and speciation events create complex evolutionary relationships between genes [2].

1.5 Post-genomic data

The advancement of experimental techniques has seen a multitude of disparate datasets. This is often confounding, as a scientist may have to access several disparate resources to ask a relatively simple question. The data available from *Salmonella* research is no exception; ranging from the gene specific through to comparison of whole genomes.

1.5.1 Metagenomics

A fairly recent branch of microbiological research is metagenomics. Research revolves around the microbial genomes collectively in a particular niche or ecosystem, the microbiome [90]. Metagenomics can give information on species and strains which are unculturable *in vitro* [91]. A study by Kurokawa *et al.* inferred that many of the micro-organisms are involved in pathways essential for host survival, which the host itself cannot carry out [91].

Microbiome members can survive within the host, carrying genes which confer defence to the host's immune responses. Due to the environmental pressure of the host the intestine is a HGT hotspot [90] meaning that pathogens can also acquire these advantageous genes.

1.5.2 Transcriptomics

The transcriptome is all of the RNA expressed at a given time. This differs from the genome in that the genome, broadly speaking doesn't change. The expression of messenger RNA (mRNA) and the other RNAs vary according to how a cell is responding to its environment, for example different genes are expressed in vivo rather than in vitro, these genes could be linked to pathogenicity [92, 93].

Microarrays are commonly used to measure the expression of mRNA or complementary DNA (cDNA). The array is comprised of a number of spots, each spot holds many copies of a unique probe. The probes are single stranded DNA sequences that can define a particular gene or region of interest. The expressed mRNA/cDNA is labelled with a fluorescent dye (this can be two dyes when looking at differential expression, for example cell infection vs. broth). The expressed strands

bind to the probes and the intensity of light produced when shone under a laser is recorded. This intensity is quantified into a measurement of concentration. The change in intensity between two samples is used to define differential expression. For example genes which show a significant decrease in expression relative to the control group are described as down regulated, conversely, those which are unregulated show a significant fold increase in expression.

Identifying differentially expressed genes can give an insight into the pathogen's mechanisms of infection and regulation. Studies have explored how *Salmonella* grow in cell culture compared to broth, this has shown what genes are required for colonisation of different cell types or hosts [94]. There has also been work looking at how *Salmonella* can infect wild type host vs. hosts mutated to be deficient in a particular gene, for example, Wright *et al.* showed that there is differential expression of genes when Salmonella infects TLR4 deficient mice [95]. Other than looking at host/environmental differences microarrays have also been used to compare wild type strains of *Salmonella* with mutated strains. These type of knockout studies can show how *Salmonella* regulates gene expression, for example IHF mutants show decreased expression of growth and virulence genes compared to the wild type [96].

It is worth noting that microarrays are not just isolated to gene expression profiling. In terms of bacterial research they can also be used for pathogen detection and SNP discovery.
The volume of gene expression data is vast, and the quality is varied. The Minimum Information About a Microarray Experiment (MIAME) standard has been developed by the microarray gene expression data (MGED) society as a basic standard for any publishable microarray experiments [97].

Two of the main public repositories are the NCBI's Gene Expression Omnibus (GEO) [98] and the European Bioinformatics Institute's (EBI) ArrayExpress [99]. ArrayExpress also has the Array Express Data Warehouse, a database of gene expression profiles from the repository which are constantly being reannotated [99].

Microarrays are extensively used but have their limitations. Each spot can be considered an autonomous experiment. However, there are many systematic errors, and normalising the data is not an easy task. The data also contains a lot of noise, and accurate measures of expression can be difficult [97, 100, 101]. Statistics and machine learning techniques (such as clustering) are often used to reduce and summarise the data.

Converse to looking at the expression of specific genes, microarray research can take a more holistic standpoint. Genomotyping is the comparison of the presence/absence of genes between genomes using microarrays [102]. Genomotyping research can help to identify genes linked to drug resistance, giving an insight into the evolution of such traits and aid classification [103]. Probes using variable genes (those varying between strains) have been developed for several purposes, including distinguishing between different strains within the same serovar [104] and for quick identification of food borne pathogens within the food industry [105]. However, genomotyping can be subject to nonspecific hybridisation; that is the probe is hybridising to the wrong area, giving a false positive. This is especially common in regions of nucleotide repeat.

In terms of NGS RNA-Seq is the transcriptomic technology, in comparison to microarrays which are limited to the spots that you have on the array, RNA-Seq can discover new genes as well as quantify existing genes. RNA-Seq is short for "RNA sequencing", though in reality most sequencing technologies actually sequence cDNA (complementary DNA). Single stranded RNA is reverse-transcribed to cDNA and the resulting pool of cDNA molecules sequenced. The number of sequences that align to a reference gene set is then used as a quantitative measure of gene expression. In eukaryotes, mRNA molecules have a polyA tail and this can be used to enrich for mRNA molecules (often those of most interest) compared to other RNA such as ribosomal RNA (rRNA) and transfer RNA (tRNA). However, in prokaryotes, mRNA molecules do not have a polyA tail, and often a ribo-reduction technique is employed to remove ribosomal RNA.

1.5.3 Proteomics

Whereas microarrays allow scientists to study mRNA abundance, proteomics allows protein abundance to be measured. Measuring protein abundance in certain conditions or over a time period, for example at different stages of infection, provides insight into how the pathogen is responding to its environment at the protein level. Finding relevant proteins can then be followed by knock-out analysis to find the effects of turning off the gene which codes for the protein [106].

One of the limitations of transcriptomics is that mRNA can be transcribed even when a protein is not being made, so the results can be misleading. Proteomics on the other hand only detects for the presence of proteins. There are two main ways of doing this they are: 2 dimensional electrophoresis followed by peptide mass fingerprinting using mass spectrometry (MS) and Shotgun proteomics which breaks the proteins into peptides, uses a chromatogram to sort by size followed by MS [107].

Proteomics can be used for different applications in bacterial research. The main use is to compare the protein profile in different states. This has been used to find novel proteins associated with infection [108]. White *et al.* used proteomic analysis to identify that periplasmic binding proteins are required to form rdar biofilm colonies [109]. It can also be used to improve genome annotation, by identifying true start sites of open reading frames and identify coding regions that were not identified via *in silico* ORF prediction [107].

1.5.4 Metabolomics

As with other -omics data, metabolomics takes a whole organism perspective, specifically of the metabolites being produced for a given state (i.e. grown on different media). Looking at this snapshot of metabolism can give insights into the pathways need for different cell states (such as drug treatments or biofilm formation) [110].

The techniques required for metabolomics are complex, sometimes requiring multiple techniques for isolating the different groups of metabolites. Different techniques have varying efficacy for extracting metabolites [111]. Once the metabolites have been extracted they are identified using spectrometry techniques such as NMR spectroscopy and MS. NMR spectroscopy doesn't require the usual complex metabolite extraction techniques, making the results far more reproducible, however MS is more sensitive and accurate [112].

Biofilm formation/regulation is often studied using metabolomic techniques. This is because many bacteria can switch between a planktonic state and biofilm formation, meaning that they carry the gene repertoire for both states. Further to this there doesn't seem to be an obvious pattern for genes linked to biofilm formation. It is believe that this is due to different combinations of pathways playing a role in biofilm formation depending on factors like growth media and organism [110]. White *et al.* 2010 used a combination of NMR and MS to look at the metabolite profile for S. *typhimurium* and a CsgD mutant which prevents the formation of the extra-cellular matrix. They were able to show that the metabolites used in the glucogenesis and major osmoprotectant pathways were up regulated compared to the mutant, suggesting the role of these pathways in biofilm formation [109, 110].

1.5.5 Mutagenesis

In terms of functional genomics a popular method of identifying genes essential for virulence in pathogens is mutagenesis. A gene is mutated and the effect on the mutant strain is observed. This can be achieved by directly enumerating the bacteria during infection studies using single strains [113] or by quantifying signals from marked strains following screening of mutants in complex pools [114]. If there are not many mutants shed by the host then the gene gets a high attenuation score; the bacterium has decreased in virulence. Mutagenesis can take a genome-wide approach, where genes are randomly mutated via transposon insertion, this is called Signature Tagged Mutagenesis (STM). Each transposon is uniquely labelled, meaning that the researcher can compare composition of input and output pools by detection of the unique signature tags' hybridisation of radio-labelled Polymerase

Chain Reaction (PCR) amplicons of the tags. This method has been used to prove that host ubiquitous strains infect each species of host with different gene sets [115].

Conversely, targeted mutagenesis looks at a particular selection of genes, isolates them and mutates them. This process is laborious compared to STM, it is however more specific and can be used to confirm the findings from STM studies [113]. Mutagenesis does have some short-comings. Firstly, some tags do not amplify from chromosomal DNA after the passage through the host. Also, some virulence genes that are knocked out may be returned as false negatives because there are other genes which have compensated for the loss of gene function.

With the advancements in NGS we see more practical applications of sequencing technologies. In 2009 Langridge *et al.* [116] developed a high-throughput sequencing (HTS) technique for genome wide mutagenesis, namely, transposon directed insertion-site sequencing (TraDIS). TraDIS creates a pool of millions of transposon mutants, a hundredfold that of STM. It has been used to successfully identify genes in Salmonella Typhi that are essential for growth in bile [116]. The potential applications of this technology are exciting but the scarcity of further publications using this method suggests that there are some limitations to the technique.

1.6 Data storage and availability

There are many diverse databases tailored to hold the data types described above. There are also multiple bacterial genome databases. This section examines some of the better known resources, assessing what they offer in comparison to the other available genome databases. Do they essentially offer the same data and if so what do they offer above and beyond the standard?

There are databases that combine genomic and post-genomic data into a central hub of information. For prokaryotes the major example of this is Xbase (encompassing ColiBase), which caters for *Escherichia coli*, *Shigella spp.* and *Salmonella spp.* [117]. The array of data is vast, including several types of genome alignment and putative orthologous genes from BLASTP searches. The database is extensively

linked to many major prokaryotic resources but doesn't hold any quantitative data [118]. However, most of the data is pre-computed meaning that the viability could be compromised if the primary data is changed; for example if a genome is reannnotated. The Pathosystems Resource Integration Center (PATRIC) on the other hand is subject to continuous update [119]. As with Colibase, PATRIC can perform multiple sequence alignments and uses the GBrowse genome browser. PATRIC also has an enteropathogenic equivalent to Pfam [62], called Enterofam. As with Pfam it uses profile Hidden Markov Models (HMMs) to identify functional families based on protein sequence [119].

There are other, smaller scale databases such as IMG [120], this has most of the features that ColiBase and PATRIC boast. It also handles fused genes, for example eutN in Typhi is recognised as a fused gene. Unfortunately, the pseudogenes are not easy to look at in the context of the genome, meaning that the eutM/N paradox might recognisable but it would be hard to see it in the light of other genomes and their pseudogenes. Finally, GeneDB is the generic database created to hold the data of small genomes. Besides the sequence data it holds extensive genomic and proteomic information integrating experimental evidence with computational analyses. It has an exhaustive array of information ranging from predicted peptide properties to expression data; which can be analysed with in-house microarray tools [121]. Its main limitation is that it contains a dearth of bacterial genomes.

The resources currently available are cumbersome in terms of their update status, mainly due to their static nature. Many are limited to one area of microbiology [122, 123] and some require a user login, which can be a hindrance when trying out different genome databases [123].The information they contain is only accurate at the time it was uploaded. If this information becomes superseded it would be a very time consuming task to make changes. Another consideration is if new data types are made, adding these to the schema could be complex. Finally, people who access these resources are limited to publicly available data, and there is no option for them to view their data alongside the data within the resource.

1.7 Webservices

The World Wide Web Consortium (W3C) defines a web service in the strictest sense, explaining that web services are the programmatic interfaces used for application to application communication via the World Wide Web [124]. Tools with a web browser interface do not fall under this classification [125]. True web services carry the disadvantage that they require some level of programming skills to utilise them. This is why web browser tools are so popular; they provide a simple way of analysing data. However, the user is still limited to a form to perform the queries, rather than the capabilities provided from web services.

Many web services are based on eXtensible Markup Language (XML), a self-describing structured text format [125].The advantage of using XML is that it is language independent and both computer and human readable.

Representational State Transfer (REST) and Simple Object Access Protocol (SOAP) communicate between the server and the client using XML. The main difference is that REST is closely affiliated with Hypertext Transfer Protocol (HTTP) whereas SOAP, rather than following HTTP, extensively uses Web Services Definition Language (WSDL). WSDL provides a model of the web service as an XML document.

Large centres such as the EBI and the University of Manchester champion the provision of web services for biological use. One example is the BioMART project which offers several different kinds of web service [126]. The simplest access is provided through a uniform resource locator (URL), where the query is defined in a URL statement. BioMart also offers SOAP and Perl application programming interface (API) access. The advantage of the BioMart project is that when a user access BioMart using the web GUI any query they perform can be translated into the URL, XML for SOAP or Perl script. This makes the transition to web services a lot easier for biologists who can see how to form relevant queries. In terms of SOAP access the EBI offers the SoapLab [127] as an example, which provides programmatic access to the EMBOSS suite.

Other international genomics centres also offer web-services. KEGG is a suite of databases for genomics, pathways and other metabolomic data. The KEGG API offers over 40 functions ranging from querying and producing coloured pathway maps to returning orthologous genes. This can all be accessed through SOAP [128].

It is worth noting that since originally writing this section KEGG have expanded their webservices to include a RESTful API and Ensembl have released REST webservice access [128, 129]. This demonstrates that webservices are still on trend and are being used by some of the more prestigious biological groups.

Classically bioinformatics tools are available as downloadable applications, where the user can access them through a graphical user interface (GUI) or the command line. With post-genomic analysis establishing itself as a major biological focus we see a transition towards web based bioinformatics services. For example, the EMBOSS suite [54] and ClustalW [55] are common applications in bioinformatics. They are typical desktop applications, and recently, both have been incorporated into a number of web services [56].

A recent database of web-services for biological sciences is BioCatalogue [130]. This resource serves as a central registry of web services, currently holding over 1000 entries. Developers are encouraged to put their service's description up and offer help to potential users. Each entry has a concise explanation and the programming methods needed to access the resource. The entry also has a list of similar resources that could be relevant to the user's enquiry.

There are projects and programs available which attempt to link web services and desktop applications into a workflow. That is the user defines a pipeline of analyses, the application then takes this and forms a workflow drawing in analyses from remote web services and desktop applications. Taverna [131] is an example of a well-established workflow application, and many web services offer Taverna access to their resources. The myExperiment project [132] works to form a community of

scientists who share their workflows rather than duplicating efforts designing workflows which already exist. Users can easily download and run the workflows locally on their computer.

Workflow applications are often installed locally and provide a means of performing a pipeline of analyses. However, clearly these web-services could also be used to provide a lightweight and flexible framework for genome annotation, displaying up-to-date and novel information from remote data sources. This would be user specific, flexible, easy to interpret and require no programming skills. This report demonstrates how such systems will be implemented to provide novel information about *Salmonella.*

As previously described in sections 1.2 and 1.5 there are many different data types available, in diverse formats. NGS data for example, is very large and it isn't economic to store this in multiple locations. NGS webservices allow the data to be located in one place and accessed when needed via the internet. Not only do webservices offer users direct access to remote data, they allow them access to tools without the problems of installation and the responsibilities that ensue such as maintenance and version control. The plug and play paradigm means that the interfaces to webservices are none language specific. The data is retrieved directly from the original source meaning that it is up to date and there is no need for regular updates.

The movement towards webservices isn't a panacea but it will allow bioinformaticians and those accessing the data to do so in a more facile way, rather than spending time on red tape such as downloading the correct module or learning a new programming language in order to use the specific tool at hand.

1.8 Concluding remarks

We have seen that Salmonella has adapted to many different niches. Some serovars are able to infect multiple hosts whilst others have very restricted host specificity often with the ability to cause severe systemic infection. Further to this, increased

virulence and host specificity are also associated with increased pseudogene formation. Of the Salmonella genomes that have been sequenced few have well defined virulence. Sequencing these could provide a link between known phenotypes and genotypes. Integrating the sequence data with post-genomic data can lead to insights in Salmonella biology. A limiting factor of many analyses that involve sequence data is annotation quality. Most annotation methods are based on homology and perpetuate misannotations when used as a reference.

The integration of genomic and post-genomic data will be performed in this project by two main methods. Namely, the analysis at genomic level integration and the development of feature specific integration software. This will be achieved by using the Salmonella model of host specificity as a basis for the integration analysis. Firstly four serovars that infect different hosts will be sequenced. The analysis stage will integrate the genomes with pathways and mutagenesis data in order to elucidate genes and pathways linked to host specificity and pathogenicity. The integration software will use web services to integrate publically available remote data and the in-house private data with genome annotation. This type of integration will help to identify misannotations and provide a platform for viewing genomic features in the context of diverse datasets.

Chapter Two

Sequencing and submission of four *Salmonella* serovars that are pathogenic to livestock

A great deal of Salmonella research focusses on the mouse model [17, 133]. Although there is merit in this, it is recognised that different serovars show different clinical signs in the same hosts [23].We have sequenced and annotated the genomes of four serovars with well-defined virulence in food-producing animals. Table 2 (Section 2.2.1) summarises these genomes and their pathogenicity in healthy outbred hosts.

The importance of these serovars is that their virulence has been described across three different hosts (namely, calves, pigs and chickens, see Table 1). The fact that we know the virulence of these strains across different hosts means by sequencing them we can link the clinical observations (phenotype) to the strains' genotypes and ultimately build hypotheses based on host specificity and pathogenicity. All of the strains were isolated from animals showing clinical signs rather than from a lab culture. This is of importance because model strains such as *S. Typhimurium* LT2 were isolated many years ago and being in a lab for such a duration of time means that the clinical signs that originally observed are not always induced upon infection. The process of passaging of many years has resulted in the loss of virulence.

This chapter describes the process of taking four raw genome sequences and their progression into fully annotated genomes residing in GenBank. The first step was to sequence and assemble the genomes, this was followed by an initial attempt at automatic annotation with some manual curation for refinement (sections 2.2.2 and 2.2.3). The stage of submission, communication and ultimately acceptance into GenBank was the bottleneck in the process, taking four rounds of corrections before the annotations were accepted. 2.2.4 states the file types needed, iterates through the resubmissions and conveys the communications with the GeneBank submission team. After successfully submitting four genomes into GenBank it became clear that there are some common pitfalls in both the submission process and automated genome annotation. Section 2.3.1 describes these, explaining the rules that slow the submission process and possible solutions for these submission hurdles. The chapter as a whole highlights that automated genome annotation isn't at a standard where a genome can be submitted to public databases like GenBank without manual intervention. The discussion 2.3.2.1 considers some of the limitations of automatic annotation. It also reflects on what is needed to make good quality genome annotations and whether groups annotating novel genomes have a responsibility to produce an annotation that goes beyond the minimum requirement to get into a sequence database like GenBank.

2.1 Aims

This chapter encompasses work published in Richardson et al 2011 [1] and Richardson and Watson 2012 [2]. The main objective is to take four serovars relevant to livestock pathogenicity from genome sequence to assembled, annotated genomes available publicly.

The aims of this chapter are:

- To annotate the assembled sequences sequenced in [1]
- To submit the annotated sequences to GenBank

- To define/discuss the rules needed to successfully submit a genome into a public databank

2.2 Methods

2.2.1 Serovars

Four servoars were isolated, namely *Salmonella* serovar Typhimurium strain 4/74 (from this point Typhimurium 4/74), *Salmonella* serovar Choleraesuis strain SCA50 (from this point Choleraesuis SCA50), *Salmonella* serovar Dublin strain SD3246 (from this point Dublin SD3246) and *Salmonella* serovar Gallinarum strain SG9 (from this point Gallinarum SG9). The four isolated serovars are summarised in Table 2. The isolation part of the process was performed by Mark Steven's group.

Serovar	Strain	Host Pathology			Notes	Evidence
		Calf	Pig	Chick		
Typhimurium	ST4/74	Acute enteritis	Acute enteritis	Asymptomatically colonises the gut	Isolated from a calf with Salmonellosis in the UK. Parent of strain SL1344	[134-138]
Choleraesuis	SCA50	Systemic infection	Systemic infection and mild enteritis	-	Isolated from a pig with swine typhoid in the UK	[137-140]
Dublin	SD3246	Severe enteritis and systemic infection	Cleared	-	Isolated from a calf with systemic salmonellosis in the UK	[138-141]
Gallinarum	SG9	Cleared	-	Systemic infection	Isolated from chicken with fowl typhoid	[139, 142]

Table 2 showing pathogenicity details for each serovar sequenced. It is worth noting that this pathology is in healthy outbred hosts.

2.2.2 Sequencing and assembly

36 cycle paired-end sequencing was carried out on an Illumina GAIIx [143], yielding between 80 and 150X coverage. SOAPdenovo [144] was used to generate *de novo* contigs, and reads aligned to a reference using Novoalign (Novocraft, Selangor, Malaysia). *S.* Typhimurium 4/74 reads were assembled on the genome and large plasmid of strain SL1344 (from this point described as Typhimurium SL1344) [145]. *S.* Choleraesuis SCSA50 reads were assembled on the genome of strain Choleraesuis SC-B67 [31] and its virulence plasmid [146] *S.* Dublin SD3246 reads were assembled on the genome of strain CT_02021853 (accession no. CP001144). *S.* Gallinarum SG9 reads were assembled on the genome of Gallinarum strain 287/91 [26]. The *de novo* and reference contigs were combined using MUMmer [147] and Gap4 [148] Assembly was carried out with help from Mick Watson.

2.2.3 Genome annotation

Sequences were annotated using a combination of GenoPipe (http://genopipe.bioinfo-portal.cdac.in/) and the pipeline described. They both work by incorporating different feature prediction software [60, 79, 149, 150]. Manual curation followed to enhance the annotation, including pseudogene assessment and assignment of start sites. This section takes the process of genome annotation described in section 1.4 and explains the design, development, and implementation of a basic prokaryotic annotation pipeline

Designing an annotation pipeline might seem obsolete, but this aspect of the project was complete prior to the other resources offering quick bacterial annotation services. This stage has still served as an opportunity to understand bacterial annotation through a hands-on approach.

The design steps of this were based on the techniques learnt on a placement at the Pathogen Sequencing Unit (PSU) at the Sanger Institute. The Sanger Institute uses a mixture of genome analysis tools, in house Perl libraries and manual annotation. At the time of writing annotation of a typical bacterial genome can take up to four

months due to the manual curation aspect, the automated step as little as 5 minutes (for a 2.5mb genome) [151].

There are five scripts/tools used in this annotation pipeline. The commands for these were put into a batch file which executes the commands sequentially.

The first command uses the gene prediction tool GLIMMER. Rather than using a reference genome to predict genes it uses an iterated approach. GLIMMER makes a training set based on long open reading frames (ORFs) which is then run through GLIMMER to predict genes. A second run of GLIMMER is then performed to increase the accuracy. This outputs a '.predict' file which is a tab delimited file showing each predicted gene and its location.

This '.predict' file is the input for the GLIMMER to tab script. This script takes the '.predict' file, parses it into Embl format, assigns a systematic ID to each gene and integrates the FASTA sequence into the parsed Embl file. The output is an Embl file with gene locations but no annotation.

The Embl file is the input for the reciprocal FASTA script, which is part of the Sanger in house library of scripts. It takes an unannotated Embl file and a reference genome as input. A FASTA search [78] is performed with every gene from the query genome against every gene in the reference genome.

```
Query_Gene1 -> Reference_Gene1
Query_Gene1 -> Reference_Gene2
Query_Gene1 -> Reference_Gene3     Match!
```

If there is a match then the FASTA is performed reciprocally. That is, the FASTA is performed with the reference gene against the query gene.

```
Reference_Gene3 -> Query_Gene1     Match!
```

If the reciprocal analysis returns a match then the annotation from the reference genome is transferred onto the query gene. Once all of the query genes have been compared to the reference genome the array is saved as Embl format.

A few modifications were added to the reciprocal FASTA script. One modification was to transfer more of the annotation across, including protein ID and annotation of pseudo-genes. The inference was also added, that is how the genes were predicted (in this case GLIMMER 3.0) and the protein ID from the gene whose annotation was transferred. The script was also modified to record any unannotated genes into an array and save them for the next stage.

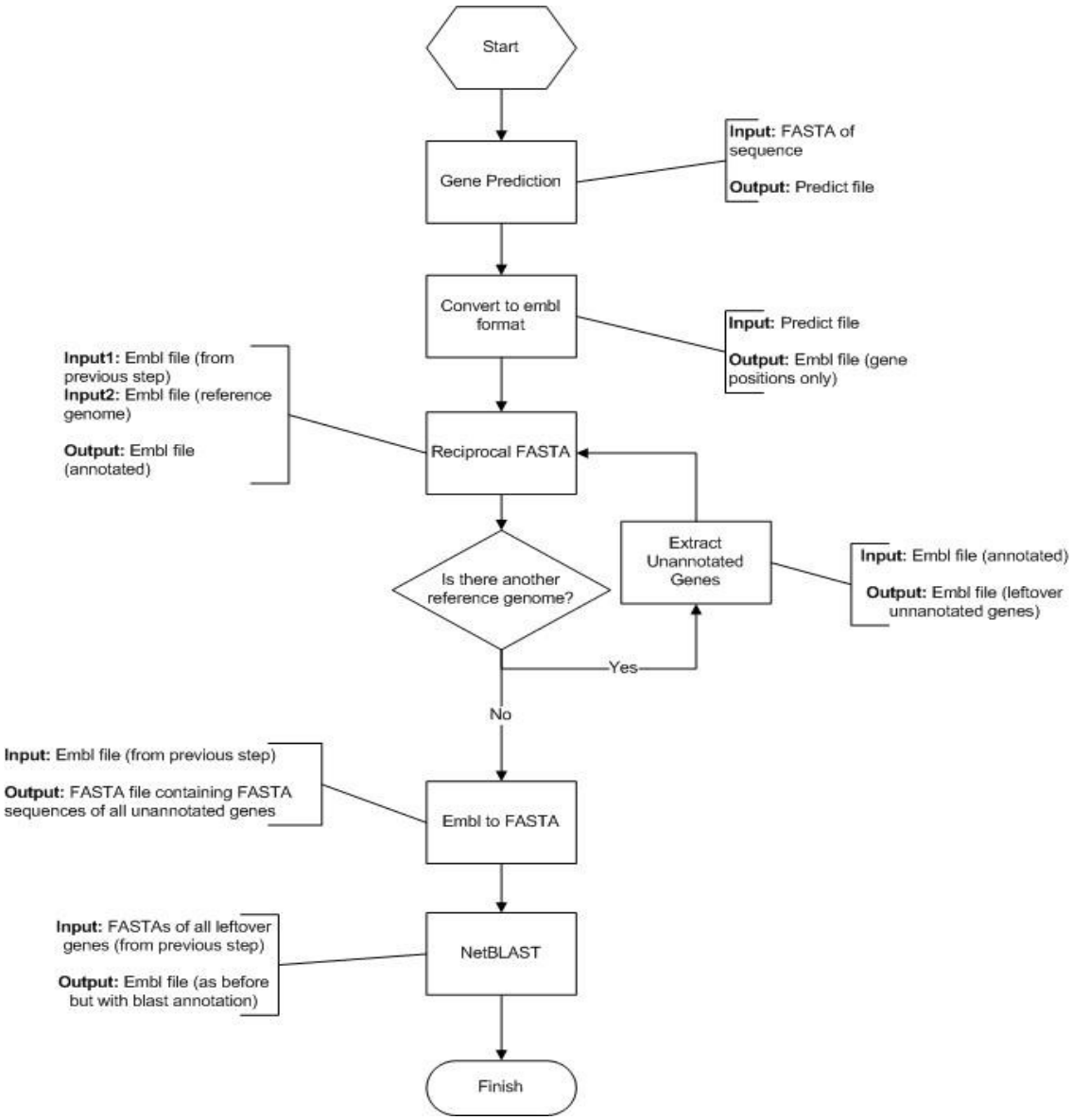


Figure 8 Flow diagram of the annotation pipeline, showing the inputs and outputs for each stage.

The next script takes the unannotated genes from the reciprocal FASTA and repeats reciprocal FASTA against a second genome of the user's choice. This script can be adjusted to run BLAST or FASTA of the unannotated genes against one of the nucleotide databases rather than against a reference genome.

Finally any genes which have not been annotated after this second run are submitted to NetBLAST (network-client BLAST) [152]. NetBLAST takes batch files of FASTA sequences and use the NCBI BLAST server to search against the user's database of choice. They are converted back to FASTA format and sent to the NetBLAST server, the results are returned in XML format. This is parsed and merged into the annotated Embl file (for the output from the pipeline see Appendix 8).

This pipeline was used for the annotation of the *Salmonella* serovars mentioned in 2.2.2. The annotation served as a basis for the manual intervention needed for submission. It takes approximately 4 hours to annotate a 5mb genome. This is slow when compared to RATT's performance but at the time of its conception was relatively fast.

As with many pipelines discussed in section 1.2.1 this pipeline does have its limitations, it will reproduce errors from homology methods. However, as the genomes will be integrated into a resource which takes remote data live from its source any mistakes should be easily detectable as they will conflict with the other data sources.

2.2.4 Submission

2.2.4.1 File formats

In order to submit the genomes to GenBank each .tbl file and the corresponding FASTA sequence were submitted to the tbl2asn software. This makes an asn.1 format file which contains the sequence data and annotation, it carries the .sqn suffix. A template file was made for each genome, using the sequin software. The template file holds the metadata (the headings in a GenBank file) such as authors and taxonomy. This file is also in the asn.1 but has a .sbt suffix.

The Typhimurium 4/74 strain was submitted with these two files. However, three of the genomes, namely Gallinarum SG9, Dublin SD3246 and Choleraesuis SCA50, could not be assembled fully, these had to be submitted as Whole Genome Shotgun (WGS) submissions. WGS submission requires .sqn files for each contig and a scaffolding file (.agp). The scaffolding file describes how the contigs are assembled and the size of the gaps between them.

2.2.4.2 Submission process

The genomes were all annotated based on genomes available in GenBank and UniprotKB. Based on this our genomes were submitted to GenBank after they were annotated according to the method in 2.2.3. The sequences and their corresponding files were uploaded using the genome submission tool [153]

These annotations were not accepted by GenBank based on various annotation discrepancies (described in 2.3.1). The volume of these was so large that scripts were needed to process the discrepancies. Figure 9shows the process for the first round of submission, the further sections explain the steps taken to make further submissions ultimately meeting GenBank's new annotation standards. A full explanation of the submission process is available in APPENDIX B.

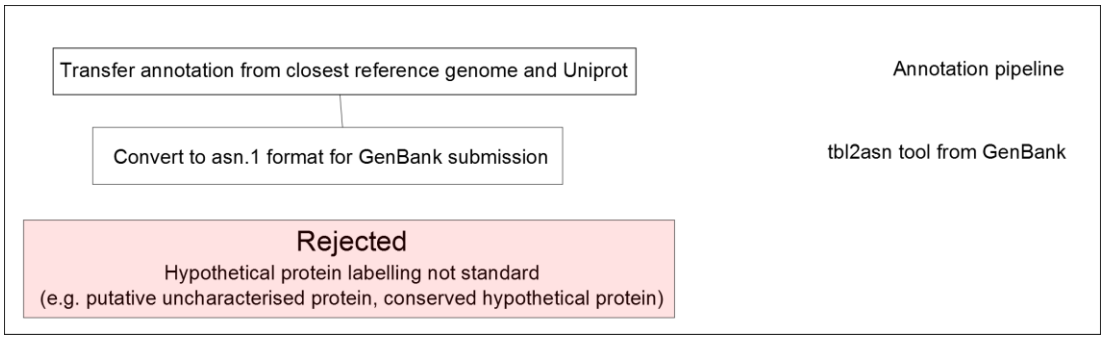


Figure 9 Schematic of the first round of submission, taking the output from the annotation pipeline described in section 2.2.3 and converting to a GenBank friendly format

2.2.4.2.1 2nd Submission

Asn2disc, the NCBI programme for detecting annotation discrepancies (http://www.ncbi.nlm.nih.gov/GenBank/asndisc.html), was run against our primary submissions and proteins without a function or those with suspect names were processed (Figure 10)

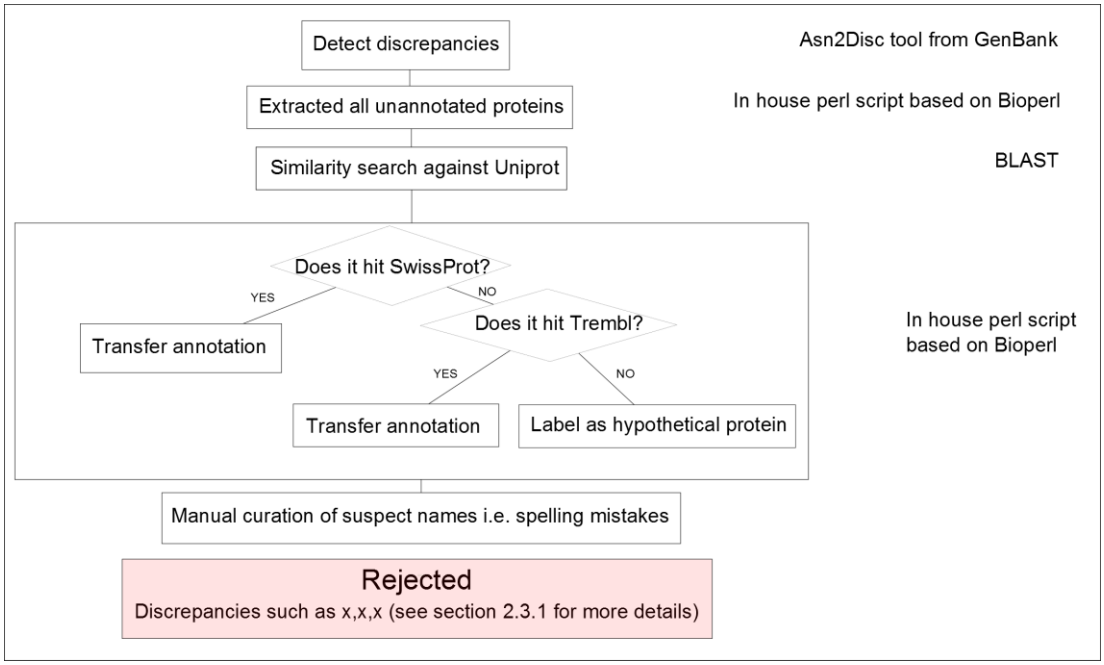


Figure 10 Schematic of the second round of submission of genome annotation to GenBank comprised of removal of spelling mistakes and reannotation proteins of unknown function.

2.2.4.2.2 3rd Submission

Many of the annotations still had suspect names according to the discrepancy file but we believed that because they hit proteins in reference genomes with functional domains that these names would be acceptable, the annotation was rejected on this basis (Figure 11).

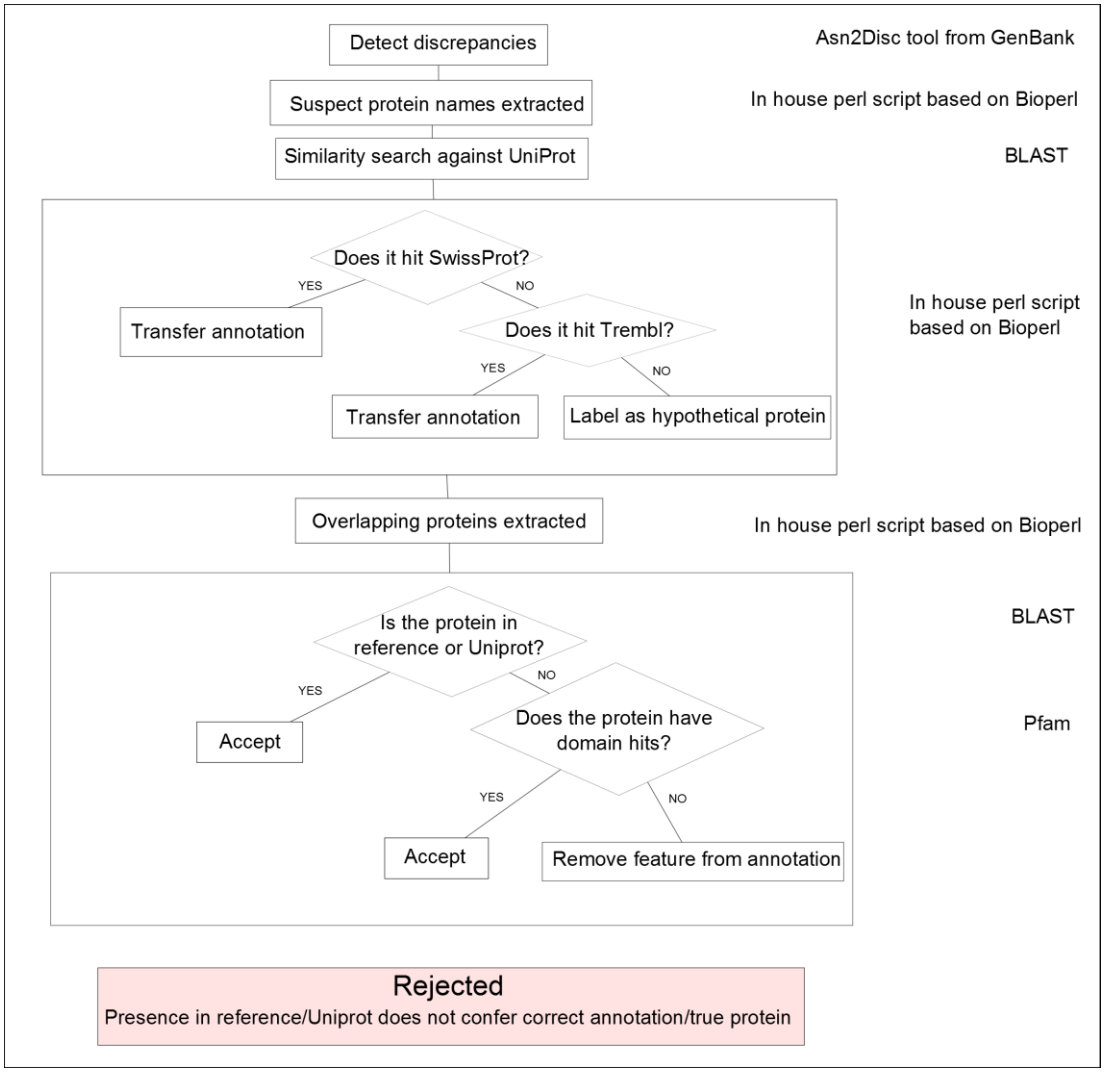


Figure 11 Schematic of the third round of submission of genome annotation to GenBank. Suspect protein names/overlapping proteins were checked against Uniprot and if present were accepted in the annotation.

2.2.4.2.3 4th Submission

All of the overlap discrepancies that were kept in the 3nd submission were present in reference genomes or UniprotKB. They were kept even if they were very short or did not have domain because of their high conservation to other genomes. Communication with the GenBank submission team explained that overlapping proteins are rare in bacterial genomes and that conservation across genomes does not indicate a true CDS, amending this resulted in acceptance of the four serovars into GenBank.

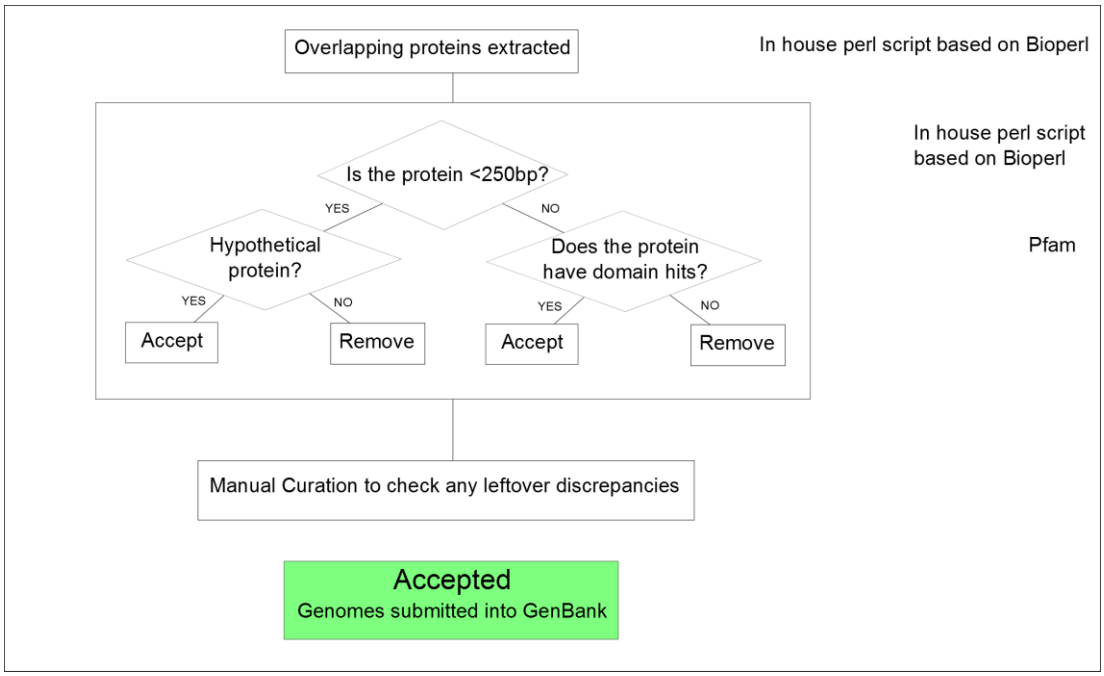


Figure 12 Schematic showing the final round of genome annotation submission to GenBank. Hypothetical proteins that were smaller than 250bp or proteins that had no domains were removed from the annotation.

2.3 Results

The genomes, their assembly details and features are summarised in Table 3. Typhimurium 4/74 has the largest genome, the most coding proteins and the fewest pseudogenes. Conversely, Gallinarum SG9 has the smallest genome, the least coding proteins and the most pseudogenes. This correlates with the gene loss hypothesis section 1.1.2.

The four serovars where successfully submitted into GenBank. Typhimurium 4/74 was the only serovar to be accepted as a complete genome. The other three serovars were submitted as WGS sequences, with the Dublin SD3246 scaffold having the most contigs. Genome maps showing the order and location of the genomes which could not be fully assembled are in Figure 13-Figure 15.

The genome annotation pipelines alone did not provide an annotation suitable for GenBank submission standards. A total of four submission attempts were required before an appropriate annotation was accepted Table 4.

Table 3 The genome summaries for each of our serovars sequenced and their plasmids

	Typhimurium 4/74				Dublin SD3246		Choleraesuis SCA50		Gallinarum SG9	
	Chromosome	TY474p1	TY474p2	TY474p3	Chromosome	p3246_74	Chromosome	pSCV50	Chromosome	Unnamed plasmid
GenBank ID	CP002487	CP002488	CP002489	CP002490	CM001151	CM001152	CM001062.1	CM001063	CM001153	CM001154
Size (bp)	5,067,451	193,842	286,908	38,688	4,842,911	74,548	4,740,379	49,558	4,658,698	87,371
Contigs	1	1	1	1	29	1	16	1	3	1
Max.contig length	-	-	-	-	1117118	-	1013670	-	2968510	-
Min. contig length	-	-	-	-	36	-	105	-	580095	-
Max. gap length	-	-	-	-	90	-	214	-	93	-
Min. gap length	-	-	-	-	1	-	2	-	1	-
Coverage	80	80	80	80	65	200	150	150	59	228
Proteins	4624	116	90	11	4580	102	4503	51	4113	86
Genes	4776	116	90	11	4819	102	4722	51	4487	87
Pseudogenes	44	0	0	0	135	0	116	0	277	1
tRNA	84	0	0	0	82	0	82	0	75	0
rRNA	22	0	0	0	22	0	21	0	22	0

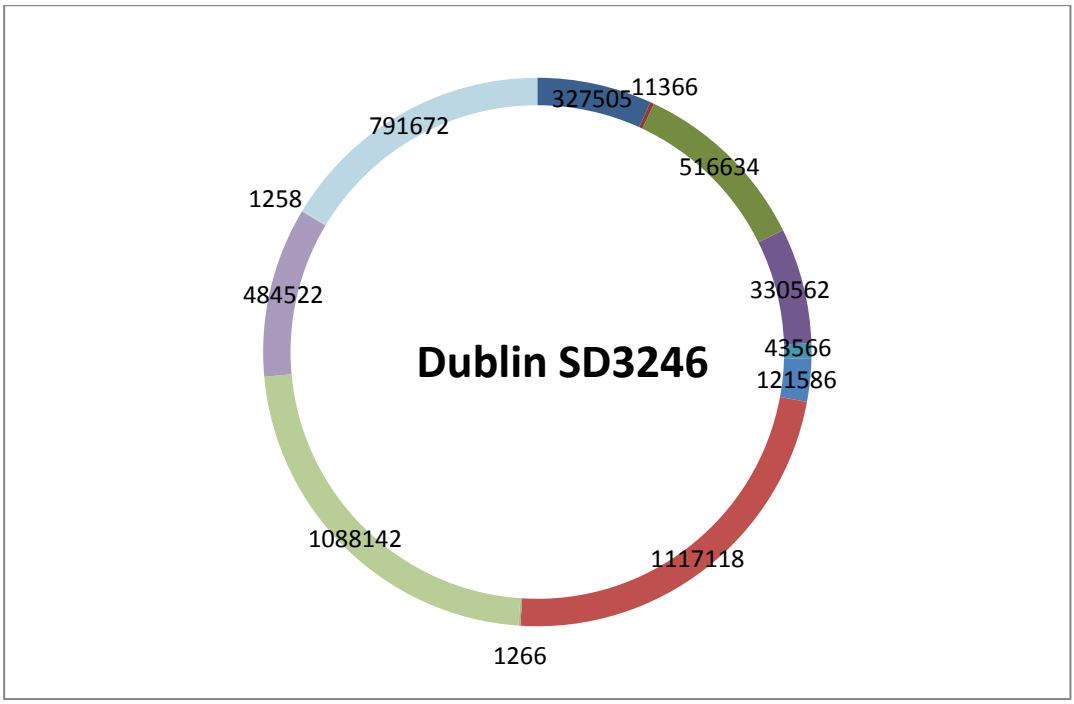


Figure 13 The de novo assembly order and sizes for the Dublin SD3246 genome (ordered against the reference CT_02021853)

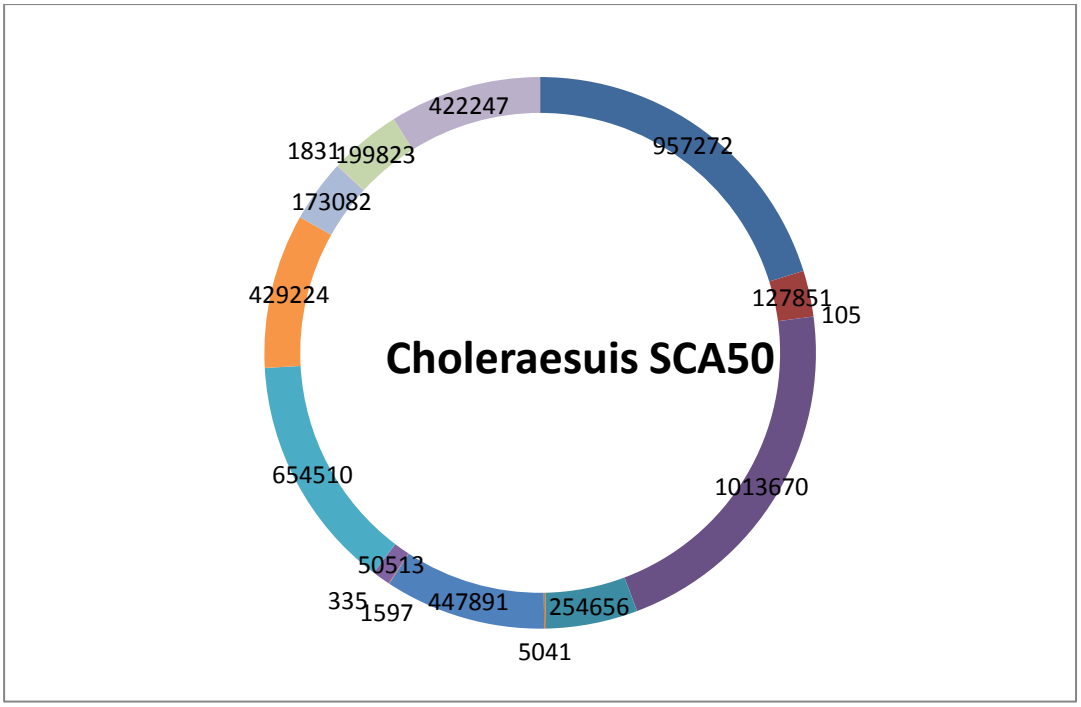


Figure 14 The de novo assembly order and sizes for the Choleraesuis SCA50 genome (ordered against the reference Choleraesuis SCB-67)

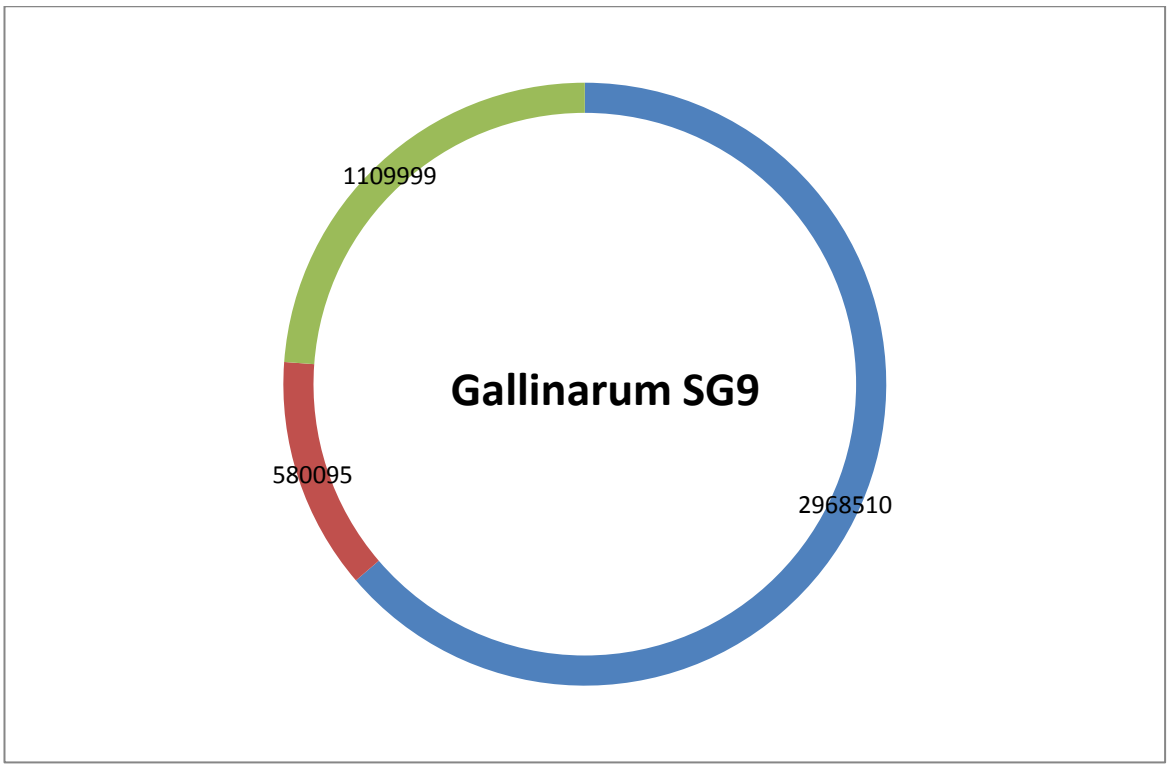


Figure 15 The de novo assembly order and sizes for the Gallinarum SG9 genome (ordered against the reference Gallinarum 287/91)

2.3.1 Validating sequencing quality

The majority of the work in chapters 2 and 3 is based on the assumption that the annotation of our sequences is correct and there are few misannotations/sequencing errors. One way of validating the quality of the sequencing is to use well known phenotypes that have been linked to the sequence and look for these in our serovars.

For example the pyridine auxotrophy of *Salmonella* serovar Dublin strains is due to a missense mutation in the nadA gene. Figure 16 shows that out serovar Dublin SD3246 holds the mutation that is unique to Dublin strains and linked to auxotrophy [154].

Our serovar Typhimurium 4/74 genome is the parent of hisG auxotroph Typhimurium SL1344. Comparisons between the two showed eight SNPs clustered around the hisG gene with one SNP actually in the hisG gene, this is expected as Typhimurium 4/74 is not auxotrophic [1].

Further to these two examples the fact that there was at least 58x coverage across our serovars is considerably better quality than the reference sequence's low coverage from Sanger sequencing. The high coverage of our data, plus the recovery of known phenotypes, gives us a measure of confidence in our sequence data and annotations. Also, the manual annotation step provided the identification of selenocysteine read-throughs in Choleraesuis SCA50 which were not detected in the original annotation of Choleraesuis SC-B67.

We were satisfied with the quality of our sequencing data due to the high coverage and linking known phenotypes to the sequence data. There are methods of systematically assessing sequence quality. For example when comparing the sequence to a reference we can identify SNPs. Randomly selecting some of these SNPs and performing PCR on these regions would give a better quality of sequence. From this we can work out the percentage of errors for the regions that underwent PCR and upscale this to the entire genome. An *in silico* based method, which

wouldn't incur the cost of extra PCR is to use something like shadow regression. This works on the premise that for any given read there are shadow reads (reads that differ by up to two bases). The number of shadow reads that are a result of sequencing error is linear to the frequency of the read and the number of legitimate shadow reads is independent of read count. A linear model is used to estimate error rates based on this [155]. We didn't incorporate any of these as we were confident that our sequencing quality was of a good standard, one reason for this is the fact that in the Typhimurium 4/74 serovar there were only eight SNPs, if there had been sequencing errors we would have expected to see a random distribution of SNPs throughout the genome (not just located around the gene known to separate it from its parent strain). If we were to repeat this work again, we would perform more stringent assessments of sequence quality.

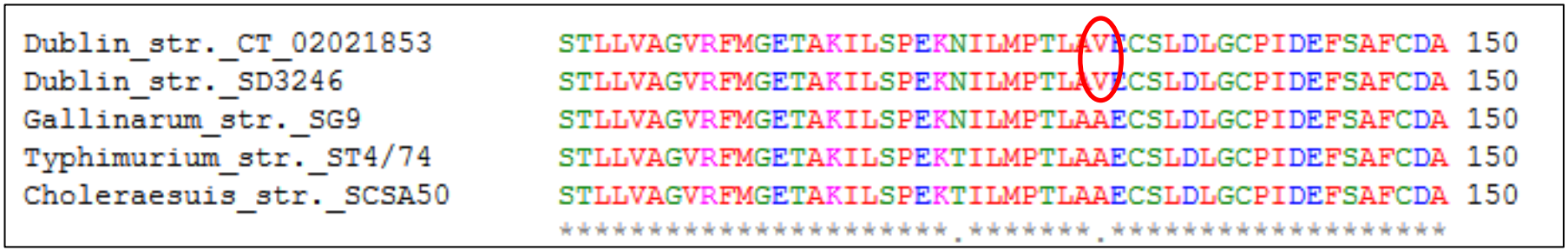


Figure 16 subsection of a ClustalW multiple alignment of nadA sequence for two Salmonella serovar Dublin strains and our other three serovars. The red circles highlights the valine/arginine difference between the Dublin strains vs. the other serovars this difference is attributed the Dublin's pyridine auxotrophy described in [154]

2.3.2 The Submission process

The process of annotation and submission has highlighted some of the common discrepancies and how to deal with them. Table 4 shows a summary of the discrepancies for each submission. Typhimurium 4/74 had the lowest number of 'suspect names', this is interesting as it had the most coding proteins. One reason for this discrepancy could be that the reference genome Typhimurium LT2 is much better studied and has a lot of experimental data to support the annotation.

Across every genome the 'BLAST hypotheticals' step increased the number of 'suspect names'. This shows that proteins flagged as hypothetical do have hits to other more informative annotations which would not be discovered by using reference genomes alone. It also demonstrates that many proteins in TREMBL (and Swissprot) do not meet GenBank's stringent annotation standards.

The 3rd submission consisted of BLAST-ing all suspect protein names against Swissprot and TREMBL. The method of prioritising Swissprot annotation and going beyond the reference genome annotation greatly reduced the number of suspect protein names. Combined with the script for reducing overlap discrepancies the manual annotation stage was more manageable, dealing with tens of discrepancies rather than hundreds (Table 4).

Table 4 Feature and discrepancy counts for each submission stage organised by serovar.

GALLINARUM SG9	1st Submission	2nd Submission		3rd Submission	4th Submission	
	Original annotation	BLAST hypotheticals	Manual Stage	BLAST suspect names	Remove overlaps	Manual Stage
Number of CDS	4145	4138	4138	4114	4114	4113
Gene Prod Conflict	47	49	49	49	49	49
Duplicate gene locus	71	77	77	77	79	79
EC number quote	1	1	1	0	0	0
Overlapping CDS	28	34	34	34	32	32
Contained CDS	42	42	42	42	12	0
RNA_CDS overlap	9	9	9	9	0	0
Suspect Product names	132	142	132	35	35	14

CHOLERAESUIS SCA50	1st Submission	2nd Submission		3rd Submission	4th Submission	
	Original annotation	BLAST hypotheticals	Manual Stage	BLAST suspect names	Remove overlaps	Manual Stage
Number of CDS	4536	4536	4536	4502	4502	4503
Gene Prod Conflict	148	152	152	144	144	142
Duplicate gene locus	216	222	222	234	234	232
EC number quote	0	0	0	0	0	0
Overlapping CDS	52	71	71	71	60	62
Contained CDS	47	47	47	47	32	30
RNA_CDS overlap	9	9	9	9	1	1
Suspect Product names	342	353	332	49	49	24

DUBLIN SD3246	1st Submission	2nd Submission		3rd Submission	4th Submission	
	Original annotation	BLAST hypotheticals	Manual Stage	BLAST suspect names	Remove overlaps	Manual Stage
Number of CDS	4626	4626	4626	4626	4580	4580
Gene Prod Conflict	54	54	54	50	50	50
Duplicate gene locus	95	99	99	99	99	99
EC number quote	0	0	0	0	0	0
Overlapping CDS	20	50	50	52	20	20
Contained CDS	38	38	38	38	0	0
RNA_CDS overlap	7	7	7	7	0	0
Suspect Product names	307	312	289	48	47	27

TYPHIMURIUM 4/74	1st Submission	2nd Submission		3rd Submission	4th Submission	
	Original annotation	BLAST hypotheticals	Manual Stage	BLAST suspect names	Remove overlaps	Manual Stage
Number of CDS	4656	4656	4656	4656	4627	4624
Gene Prod Conflict	60	61	61	57	57	57
Duplicate gene locus	99	99	99	101	101	101
EC number quote	0	0	0	0	0	0
Overlapping CDS	88	93	93	91	86	86
Contained CDS	32	32	32	32	0	0
RNA_CDS overlap	9	9	9	9	1	1
Suspect Product names	87	92	81	27	26	15

2.3.2.1 Examples of acceptable protein names that are flagged as suspect

The program Asn2Disc gave a detailed description of potential annotation discrepancies. However, some features labelled as discrepancies, although correctly annotated are exceptions to the general GenBank annotation rules.

For example STM474_1118 is labelled as ***'bifunctional protein putA'.*** The label ***'bifunctional protein'*** should not be accepted but it as it is a hit to a Swissprot protein this was accepted. Proteins with the word gene will flag up a discrepancy too, however, the protein name ***'gene 60 protein'*** is allowed for submission.

Some CDSs overlap one another by a small number of bases. This is acceptable but flagged up by asn2disc if both features have the same product name, this can be indicative of a pseudogene or split gene. However, some product names are so generic that many features in the genome will share the same product name, and some of those will reside next to one another. Descriptions such as ***'hypothetical protein'***, ***putative exported protein'***, ***'tail assembly protein'*** and ***'putative membrane protein'*** are all acceptable overlapping descriptions.

2.3.2.2 Problems with coding regions

The NCBI validation software flags up all instances where a coding region completely contains another coding region on the opposite strand. The submitter is asked to check these coding regions and decide whether these are true features. If the coding region only hits hypothetical proteins and doesn't contain any domains, it may be either removed or demoted to a miscellaneous feature.

2.3.2.3 'Same gene name, different product name'

This issue occurs when two features, either within or between genomes, are assigned the same short gene name yet different product names. The NCBI validation software specifically highlights when this occurs intra-genomically with the description 'Same gene name, different product name' [82, 83].

2.3.2.4 CDS Nomenclature

There are many words which may be unacceptable in protein names, such as ***'binding'***, ***'domain'***, ***'like'***, ***'motif'***, ***'gene'*** and ***'homolog'***. Submitters may be encouraged to change these: for example ***'bacteriophage replication gene'*** can be changed to ***'bacteriophage replication protein'*** and ***'peptidyl-tRNA hydrolase domain protein***' can be changed to ***'peptidyl-tRNA hydrolase protein'***; a note may be added to state that the feature contains the aforementioned domain.

2.4 Discussion

Although the methods in this chapter may seem straightforward, the entire process took a lot longer than anticipated. This was mainly due to higher standards of annotation required for new genome submission compared to the currently available genomes. This skew in annotation quality makes it difficult to produce a good quality submission when most of the available genomes are substandard in terms of GenBank criterion.

Genome annotation does predominantly rely on orthology to other sequences (e.g. genomic, protein, domain). When a similar high quality reference is available the annotation is quite straightforward. This is apparent when comparing the number of discrepancies in Table 3. Typhimurium 4/74 took most of its annotation from a better quality annotation compared to the other genomes.

Further to the quality of reference genomes the orthology transfer has other limitations. If the reference genome carries a pseudogene, where the query genome does not, the annotation could be completely missed or misannotated as a pseudogene.

2.4.1 The submission process

The process of successfully submitting to GenBank took a long time. This was mainly due to the time between communications with the GenBank submission team. For example the overlapping proteins that were kept during the early submissions were only brought up as a problem in after the third submission. Rather than

explaining all the problems with our annotations we were shown new problems with each submission (which we had assumed were up to submission standard because they weren't highlighted as a problem originally). If they had given all the feedback in one email it could be a bit disheartening for the annotator but in terms of time taken for a successful submission, this would be greatly reduced for both the submitters and the GenBank team.

The format for submission is ASN.1, this is complex, it is not human-readable, it is very bulky and requires special tools for conversion. It would be difficult to make a tab-delimited format of an annotation submission because each feature has many different tags (potentially meaning a lot of empty columns). XML format allows for a more straightforward, human-readable format, which would be easily parsed and edited during the annotation process, rather than converting between multiple formats.

2.4.2 Improving automated annotation

Advances in sequencing technologies are allowing researchers to sequence microbial genomes at a huge rate. It is becoming harder to devote time to manually annotate these genomes, leading to a rise in automatic annotation pipelines. However, due to a range of problems, the output of these automatic annotation pipelines is unsuitable for publication. Some changes can be made to improve this output; however, it is difficult to envisage an end to manual checking and curation.

2.4.2.1 Spelling mistakes

There are 128 proteins in UniProt that contain the word ***'syntase'***, an incorrect spelling of the word ***'synthase'***. To put this into context, the RefSeq entry for *Rhizobium etli CFN 42* (accession NC_007761) assigns the function ***'dihydrofolate syntase'*** to gene folC. This has propagated into other databases such as UniProt (accession: Q2KE79), KEGG (accession: RHE_CH00024), and xBASE (accession: RHE_CH00024). If a user was to visit any of these databases and search for ***'dihydrofolate synthase'*** the misspelled entries would be omitted from the search results. Large scale detection and correction of spelling mistakes in public databases is a difficult task, and so there is a reliance

on the submitter to correct these. Automatic annotation pipelines simply copy and propagate what is there already. Spelling mistakes may be highlighted by the validation software provided by the public databases during submission, however, an alternative correct spelling isn't offered, making it difficult to amend the mistakes without manual intervention.

This can be solved by writing rules to find spelling mistakes [76]. However, this approach is limited to spelling mistakes which are explicitly written in the code. A solution may exist beyond biological science. The search engine Google upon receiving the input 'syntase' automatically states 'Did you mean: *synthase*'. There are programming languages which have classes or plugins to produce such 'did you mean' results [156, 157].

2.4.2.2 'Same gene name, different product name'

In the current set of 2696 microbial genome and plasmid sequences in RefSeq, we detected 23,843 genes with at least two different product names (see http://www.ark-genomics.org/genomeannotation.html for the full list). The most extreme example of this is gene 'tnp' which has 151 different product names ('tnpA' has a further 97). A more manageable example can be seen in Table 5. The 'int' gene has a total of 12 different product names across 17 *Salmonella* RefSeq entries. These product names contain huge variation in terms of information content. When using an automatic annotation pipeline, there is a danger that if the top hit is to an entry labelled ***'Hypothetical protein'***, then you will capture far less information than if your top hit is ***to 'phage integrase family site specific recombinase'***. In order to correctly annotate this gene in a new genome, it is necessary to take into account all of these product names in the annotation process. It is difficult to imagine a set of text-mining rules that could efficiently interpret the range of annotations and assign the most suitable one to a new gene.

Table 5 Different product names assigned to features with the gene name '*int*' across 17 different RefSeq entries for *Salmonella* species [2]

Gene name	Product name	Accession
int	bacteriophage integrase	NC_003198, NC_004631, NC_015761
int	Gifsy-1 prophage Int	NC_006905
int	hypothetical protein	NC_006905
int	Integrase	NC_003198, NC_004631, NC_006511, NC_012125
int	integrase (fragment)	NC_003198
int	phage integrase family site specific recombinase	NC_006905
int	putative cytoplasmic protein	NC_006905
Int	Putative integrase	NC_003384
int	putative integrase protein	NC_006905
int	putative P4-type integrase	NC_006905
int	putative phage integrase protein	NC_006905
int	site-specific recombinase, phage integrase family	NC_012125

2.4.2.3 Hypothetical proteins

The term **`'hypothetical protein'`** often refers to a gene that has been predicted by software but which finds no homolog of known function in the databases, and which has no known functional domain. There are currently 53 035 proteins whose product name contains both words in UniProt (search term: 'name:hypothetical AND name:protein') and there are a further 5 178 212 proteins in UniProt that contain the words 'uncharacterized' and 'protein' (search term: 'name:uncharacterized AND name:protein'). These may be real genes with no known function or they may be artefacts of the gene prediction process.

Many bacterial genes of unknown function are assigned y-gene names based on their orthologous location in *E. coli K*-12 [158]. The letters denote the location in terms of minutes around a circular genome. This gene annotation has propagated throughout many strains and species of bacteria, losing the relevance and context of its name as the genes are not all in the same relative location to the original annotation in *E. coli K-12*. For example the *yabF* gene has a known function, 'glutathione-regulated potassium-eflux system ancillary protein'. The gene name *yabF* is completely meaningless in all genomes other than the original and actually has a synonym *kefF*. With that in mind annotators should use more informative gene names as a preference, choosing alternative gene names over the original y-gene annotation.

Often there are features which are only orthologous to other hypothetical features and do not contain any domains. These could either be regions with no functionality, a relic of the feature prediction software or the domains present have not been discovered yet. Whether or not to include them is often a decision made by the annotation team and varies between groups. Thus, many artifactual 'hypothetical proteins' may be annotated, published and disseminated into the public databases, reinforcing the annotator's belief that their new gene predictions do indeed have homologs in other species. It would be more informative to actually state in the annotation a score for each feature. This will allow users to make informed assessments of the features and programmers to easily parse genomes to handle

hypothetical proteins based on their quality of hits. Gilks *et al.* [159] discuss the possibility of assigning scores based on the source of annotation.

There are arguments for and against keeping these proteins in the annotation. If they are indeed a misannotation by the gene prediction software they should be removed as they will perpetuate through secondary and tertiary databases as a recognized protein awaiting functional discovery. Searching for conserved domains or motifs in databases such as Pfam or InterPro can give an indication of whether a hypothetical protein is functional but this has pitfalls too. The fact that a protein has a domain hit doesn't necessarily convey its function. Pfam [160], for example, contains over 3000 "domains of unknown function", or DUFs, representing over 20% of known families [161] and as more novel genomes are sequenced the number of new DUFs will increase. A hit to a DUF does not inform us of a feature's function, but as they are areas of high conservation they indicate a potential region of biological interest.

Through computational methods alone there are no means to conclusively determine whether a genomic region is functional. With that in mind conserved features of unknown function should be kept because in the future they may be recognized as a true region of interest; however, they should be annotated differently to discriminate them from features with stronger evidence. Evidence tags are available but they are often not present, and are not a prerequisite for submission to GenBank or Embl. Evidence qualifiers such as how the feature was predicted (e.g. GLIMMER, BLAST, homology) and what entries it hits in a given database provide a clear audit trail for anyone who wants to assess the quality of a particular annotation. The type of data source used, that is, whether it is manually curated or automated should be stated, providing the user with a clear method of judging the annotation. As experimental data becomes more ubiquitous evidence tags should play a larger role in annotation.
Experimental methods such as RNA-Seq [162] and Signature Tagged Mutagenesis (STM) [163] may help to identify regions of functionality. RNA-Seq data can help delineate and quantify areas of transcription, and overlaying this expression data on the genome may help biologists to identify pseudogenes and the true locations of

features. STM can help identify the function of genes by monitoring the phenotype of single-gene mutants.

The most important point is that one's annotation is only ever as good as the reference data sources. In terms of publicly available genome sequences the quality is varied. It is worth actually looking at the annotation and assessing the quality. Choosing a genome because it is the closest relative will give the most homologous features but might not give the best quality annotation.

Combining additional data with the original annotation gives scientists a new way of viewing the genome. Experimental data could be able to solve the *eutM/eutN* problem described above; for example, RNA-Seq data would show which areas of the genome are actively transcribed and STM may indicate whether knocking out either of the genes alters the phenotype of the mutant.

Many scientists go through the process of annotation with the final aim of submitting to a genome database such as GenBank or EMBL. In order to realize this goal there are many rules which need to be followed [82, 83] and often validation software is provided to verify one's annotation. These rules are imposed to ensure a better standard of genome annotation, however, they do mean that often the output of an automatic annotation pipeline must be manually checked and altered prior to publication. Many of the issues described in the 'Limitations of the Annotation Process' section may be identified as potential problems and the submitter is provided with long lists of features that represent these. They must be checked, and either altered or justified. In addition to those mentioned above, there are others described below.

2.4.2.4 CDS nomenclature

These rules add complications if the submitter wants to fully automate the process of annotation. As a rule of thumb, if a predicted coding region has homologs in SwissProt these are the best protein names to transfer across and running the validation software after using SwissProt initially can greatly reduce the number of

suspect names. As an aside, ***'probable'*** and ***'predicted'*** are not flagged up by the validation software but ***'putative'*** is the preferred alternative.

Some CDSs have the same protein name as the protein next to them, which can be the sign of either a disrupted gene or a valid gene duplication event. It can also be because the protein name is very general such as ***'hypothetical protein'*** or ***'inner membrane protein'***. These features may be flagged up by the validation software and, if they are not pseudogenes, need a note stating that they overlap a CDS with the same protein name.
CDS gene names that appear more than once in a genome and have different proteins names to one another (e.g. Table 5) may also be identified as potential errors. These may be brought to the submitter's attention who often has to use their discretion and knowledge to assign gene names correctly. This can be as simple as performing a similarity search and seeing which gene names are associated with the hits.

2.4.3 Gold standard genomes

RefSeq is one attempt to standardize and improve the quality of genome annotation; however, as we have shown, problems persist. With the implementation of stricter rules for submission we should see an increase in annotation quality. While genomes of varying quality are available there should be a means for scientists to see the quality of any given annotation. Evidence qualifiers such as how the feature was predicted and what entries in a given database the feature sequence hit, including the database version and date, would provide a clear audit trail for anyone who wants to assess the quality of a particular annotation. The type of data source used, that is, whether it is manually curated or automatically generated should also be stated, providing the user with a clear method of judging the annotation.

Out of the 1851 publicly available completed bacterial genomes 102 have a version number of .2 or higher [164]. This means that the submitting group have revisited the original sequence and changed it. The fact that the sequences have been changed is indicative of a higher quality sequence. This, however, does not reflect the quality of the annotation. It is possible to look at the revision history of genomes within GenBank, this will give users an idea of changes on a genome by genome basis, no

small feat when there are 1851 genomes available. In the literature there have been several papers which have revisited and reannotated genomes, these include strains of *E. coli, Campylobacter jejuni* and *Mycobacterium tuberculosis* [165-167]. In terms of what is currently available these genomes are likely to be the closest to realizing 'gold standard genome annotation'.

Janssen *et al.* [168] calculated the number of publications per gene for all completed genome to calculate a Species Knowledge Index (SKI) for each genome. They showed that, in bacteria, there is a pronounced bias toward certain organisms namely *E. coli*, *Pseudomonas aeruginosa* and *Bacillus subtilis*. With this in mind perhaps there should be a focus to annotate genomes with a high SKI to the highest level possible as there is such an abundance of experimental data available. These can then be used as gold standard genomes for annotations of other species.

As we learn more about genes and protein function it becomes clear that a simple protein name is inadequate. Some proteins are multi-functional, performing different tasks depending on the context it is expressed in. We can say that a protein has a one-to-many relationship with function, meaning that assigning a protein name based on the first function associated with it can be misleading and inaccurate. The Gene Ontology (GO) may provide a more flexible way of describing a range of functions explicitly and concisely, and GO annotations natively include evidence qualifiers. However, GO terms are not frequently included as part of the initial annotation of bacterial genomes. The EBI offer UniProtKB-GOA Proteome Sets [169], GO annotations for all completely sequenced genomes in the public domain, however, these are not included with or clearly linked to the original genome submission. The development and use of GO annotations is encouraged and these should be included in genome annotation efforts.

2.4.4 Going beyond the minimum

For acceptance to databases such as GenBank or EMBL, only gene, CDS and structural RNA features need to be added [82, 83]. However, many other features should be added. This section gives a broad overview of some of the other features and how they can be predicted; a comprehensive guide is available [170].

Gene prediction software sometimes assigns the wrong start/termination sites. GLIMMER for example assigns the start site as the most upstream start codon [171]. By searching for ribosomal binding sites (RBS), one can infer and reassign the start site; RBSFinder does this by looking for motifs such as the Shine-Dalgarno sequence pattern [171]. For termination sites, TransTerm searches for rho-independent transcription terminators to assign the correct termination site [172]. As well as correcting start/termination sites these features should be added to the annotation, using the tags 'RBS' and 'terminator' respectively.

Regions of conservation within proteins such as motifs and domains should be added to the annotation after the gene finding step. There are many databases which store protein families such as ProSite, PRINTS and Pfam [160, 173, 174]. InterproScan can perform searches against a range of domain/motif databases [175]. Hits to motif/domain databases should be assigned the qualifier 'db_xref' within the corresponding CDS feature [82, 83]

Areas of horizontal gene transfer (HGT) such as pathogenicity islands and prophage can be predicted by looking at asymmetries in codon composition and the GC content as these will often differ between areas of HGT and the rest of the genome [176]. They are often associated with the presence of integrases, transposases and IS elements [176]. Software tools exists to predict these [21, 177], and these are reviewed and compared in by Langille *et al.* [178]. There are clear guidelines for annotating phage, this should be assigned under the 'source' feature with the name of the bacteriophage in the 'organism' qualifier and the type of sequence in 'mol_type' (usually genomic DNA). There is no specific annotation tag for other GIs so these should be annotated as miscellaneous features. The mobile genetic elements themselves use the 'mobile_element' tag.

Sequence repeats such as "clustered regularly interspaced short palindromic repeats" (CRISPRs) and other tandem repeats are of biological interest. For example, they can be used to understand the bacterial defence mechanism [179] and to distinguish between closely related strains [180]. Software tools exist [181, 182] and databases

such as MICdb store predicted microsatellites as well as offering a prediction tool for user inputted sequence [183].

Identifying a protein's cellular localization can be indicative of function and this can be used in the identification of drug targets. There are many methods of prediction including homology and keywords [184], amino acid composition [185-187] and a mixture of these [188], Gardy *et al.* have performed a comprehensive review of the many tools available [189].

2.5 Concluding remarks

The entire procedure took several months, primarily due to two post-assembly bottlenecks. Firstly, a manual annotation stage was required for assessment of features such as pseudogenes. Secondly, the requirements for a successful submission into GenBank are more stringent than they were in the past. This became evident when some annotations transferred from genomes already in GenBank were not acceptable. The fact that GenBank has made a move towards higher quality genomes is commendable, it does however slow down the submission process as the 'old' genomes, which are commonly used for annotation transfer, are not up to the more stringent annotation standard.

The pipelines currently on offer do not take many of the pitfalls outlined above into account, meaning that a lot of manual effort is required to correct errors and inconsistencies. It is easy to imagine adjustments to current pipelines that take into account certain aspects (e.g. common spelling mistakes) but not others (e.g. correctly interpreting pseudogenes). Realistically, completely removing the manual stage of annotation would be imprudent, however, improving current automated pipelines may greatly reduce the time spent manually checking the annotation. There have been a flood of new genome-wide data types in the post-genomic era, for example microarray and RNA-Seq data, many of which can assist with genome annotation. However, these are often large, unwieldy, come in a variety of different formats and can be hard to integrate with one another. Allowing scientists to visualize this data alongside genome annotation can be hugely powerful [190]; however, genome annotation is often kept in specific flat file formats where integrating non-text data is

virtually impossible. Secondary and tertiary databases may include additional data alongside the original genome annotation [20], but these "data warehouse" approaches employ copies of the original data which can become out-of-date and out-of-sync with the original data. The advent of bioinformatics web-services [191] may allow new systems that query data live over the internet, ensuring the latest data is displayed.

If we had to do this work again it would take considerably less time. The time and cost of sequencing is reduced compared to two years ago when we started the process. The software available for de novo assembly is now faster and more accurate [192, 193]. Further to this read lengths are becoming longer (from 36bp to 250bp on Illumina Miseq). With this in mind it is likely that we would have been able to submit all the genomes as complete rather than WGS. Also, there are annotation tools that claim to take less than four minutes to transfer the annotation across to closely related bacteria (compared to the 4 days it originally took) [151]. Now that we are familiar with the GenBank annotation requirements means that less time would be spent revising resubmissions and some scripts could be written to automate the process further.

There is a need for reannotation of the currently available genomes as these are often used as reference genomes. This project did not add to the effort of reannotation of reference genomes, this is predominantly due to the amount of time to get four novel genomes accepted. In terms of future work, with the experience that we gained in the submission process it is feasible that we could add to the reannotation effort. I believe that the main factor that will contribute to improving reannotation is the inclusion of trackable trail of annotation. Meaning that it would work like citations in a scientific paper, one can ultimately reach the primary source of the original research. For reannotation, specifically genomes that are used as model annotations/organisms the inclusion of experimental evidence from the literature would be a good starting point to validate annotation and remove hypothetical proteins. Mining the literature for this kind is not a small feat, semantic literature techniques are available but are not commonly used. Efforts like GeneDB have

heavily manually annotated genomes, until we get to a standard which allows for the integration of diverse experimental results into the annotation we will have to continue with manual annotation as the highest quality method of annotation.

The fact that the virulence of these strains across different hosts is well defined means that the genomic data can be used to make hypotheses based on host specificity and pathogenicity of the serovars in different hosts. Additional data from post-genomics experiments can help improve genome annotation; however, a line has to be drawn regarding what data should be included in the annotation and what should be in separate databases. Tools and services need to be developed which offer scientists a means of viewing genome annotation augmented with other experimental data. This will empower the user to make meaningful judgments on the quality of annotation and the relevance of a particular region to their research.

For the foreseeable future bacterial annotation requires both automated and manual steps. Offering users a measure of quality for the whole genome and individual genes will allow user to make an informed choice regarding reference genomes and transferring annotation between genomes. Using GO terms, for example, would improve protein description and reduce syntactic errors.

Chapter Three

Functional analysis of *Salmonella* genomes for signatures of host specificity and pathogenicity.

Broadly speaking the relationship between host specificity and pathogenicity is well defined in *Salmonella*, with the majority of strains adhering to the model described in 1.1.2. The mechanisms behind these phenotypes are still being explored. The sequencing and annotation of the four servoars described in chapter 2 are used in this chapter as a basis for investigating the patterns between host specificity and method of infection. Our serovars are of particular interest because they have well defined virulence due to the fact that they were isolated from infected livestock.

The chapter can be divided into two main sections. The first part, 3.2, links KEGG pathways to the annotations and explores patterns of pseudogene formation in these in an attempt to answer the following questions:

- Do strains in the same serovar show similar patterns of pathway attrition and how suitable are our strains as representatives of their serovar? (3.2.2.1)
- Is pseudogene formation in KEGG pathways random? (3.2.2.2)
- Are there any pathways that only show pseudogene formation in one serovar? (3.2.2.2.1 – 3.2.2.2.4)

- Are there any pathways that show pseudogene formation across all servoars except one? (3.2.2.2.1 and 3.2.2.2.4)
- What pathways are enriched for pseudogene formation in each servoar, showing more pseudogene formation that expected by random chance? (3.2.2.2.1 – 3.2.2.2.4)

Section 3.2.3 tries to apply some biological interpretation to the results such as the use of one strain as a representative for its serovar and whether the suitability is actually linked to host specificity. A major finding was that pathways linked to genetic processing such as 'nucleotide metabolism' and 'replication and repair' were highly conserved across all serovars, conversely the variation between servoars was predominantly seen in metabolic pathways (section 3.2.3.2).

The other main section of this chapter, 3.2.3.5, uses genomic wide mutagenesis (TraDIS analysis) of *Salmonella* serovar Typhimurium strain SL1344 across chicks, calves and pigs to firstly look for areas of negative selection compared to pseudogene formation in serovars of known pathogenicity. Further, the genes in the TraDIS data were assigned to pathways and enrichment analysis was used to elucidate pathways that show more positive/negative selection than random chance.

3.1 Aims

The purpose of this chapter is to augment the genome annotations from chapter two with post genomic data. Showing that by integrating functional descriptions and experimental data one can elucidate patterns between genotypic information such as pseudogene formation and phenotypes like host specificity and pathogenicity.

The specific aims of this chapter are:

- To assign KEGG pathways to gene/pseudogene features
- To compare areas of pseudogene formation within serovar to assess the use of serovar representatives
- To compare areas of pseudogene formation between serovars
- To find enriched pathways for negative selection in the TraDIS data
- To assign orthologs from our serovars to the TraDIS data
- To use the ortholog assignment to identify areas of gene loss compared to the TraDIS data

3.2 Pseudogene Analysis

KEGG is a resource comprising of three areas namely; genomic, chemical and network information. Within the network section lies the pathway database, this consists of manually drawn maps for metabolic pathways [64].

This chapter refers to KEGG pathways in many of the figures. KEGG describe their pathway maps as "molecular interaction/reaction network diagram represented in terms of the KEGG Orthology (KO) groups, so that experimental evidence in specific organisms can be generalized to other organisms through genomic information" [64], for example Figure 27 shows the KEGG pathway for Starch and Sucrose metabolism. There are three main types of object which are explained below (from the KEGG description [64]):

- boxes - genes or gene products identified by the combination of the KEGG organism code and gene identifiers
- circles - other molecules, usually chemical compounds identified by C numbers, but including glycans identified by G numbers
- lines - reactions identified by R numbers in metabolic maps; ortholog (KO) groups identified by K numbers in global metabolism maps

These maps can be integrated with individual genomes using the EC numbers assigned to the genomes' genes [64]. The pathways are organised hierarchically with pathways and higher descriptions (the full pathway map is available Appendix A: KEGG_pathwaymap.docx).

The relationship between pathways and genes is 'many to many'; this means that each pathway can be assigned to no genes, one gene or multiple genes. Conversely, each gene might be associated with no pathways, one pathway or many pathways. Assigning genes to their respective pathway is a useful method of grouping genes, based on their function. Although pseudogenes aren't technically part of a pathway, they are assigned a pathway based on the original gene function.

In this chapter pathways which had pseudogenes assigned to them are referred to 'pseudogene pathways'. Conversely, pathways which only had non-pseudogenes assigned to them are described as 'functional pathways'.

This section describes the assignment of KEGG pathways to the genes and pseudogenes in each serovar sequenced in Chapter 2. Once assigned, a functional enrichment test can be used to look for pathways which have significant gene loss in each serovar. This can give an insight into which pathways are essential/non-essential for each serovar potentially implying the pathways required for infecting different hosts. Looking at pseudogene formation in pathways can give a different perspective to the host specificity/gene loss model for example two servoars that share a similar phenotype but don't have the same pseudogenes might share pseudogene formation in the same pathways. The concept of this type of resilience in biological pathways is discussed in section 3.2.3.

The analysis of pseudogene enrichment in KEGG pathways provides a solid foundation for making testable hypotheses around the *Salmonella* host specificity model. However, KEGG pathways are bite size chunks of a single network, the pathways are all interlinked. In order to remove the bias associated with KEGG pathways pseudogene formation should also be looked at in the context of the entire KEGG network. This can be achieved by clustering the network and looking for clusters that are enriched for pseudogene/gene absence. This type of analysis is complicated by the fact that KEGG now require a subscription fee for ftp access, running software that can handle the entire KEGG network requires a considerable amount of computing power and the analysis itself requires a considerable amount of time, this section provides a proof of concept, exploring how this type of analysis could be performed and some of the difficulties associated with it. The methods are described in 3.2.1.5 and as this is a proof of concept and there are no results the outcome of this section is discussed in 3.2.3.5.

It is worth noting that at the time of writing this the serovars were in the process of submission and there was no automatic pathway assignment available in KEGG.

3.2.1 Methods

3.2.1.1 Intraserovar Analysis

For Gallinarum our strain SG9 was compared to the only publically available *Salmonella* serovar Gallinarum strain 287/91 (from this point Gallinarum 287/91). Our Choleraesuis strain, SCA50 was compared to *Salmonella* serovar Choleraesuis strain SCB-67 (from this point Choleraesuis SCB-67).

The pseudogenes for each strain were extracted from their annotation using the ***get_pseudogenes.pl*** script (Appendix A: get_pseudogenes). A file of FASTA sequences was made for each strain using ***embl2FASTA_leftovers.pl*** (Appendix A: embl2FASTA_leftovers.pl).

These FASTA sequences were BLAST against a local copy of the KEGG database using BLASTn. The results were parsed and the pathway counts were calculated using ***BLAST_kegg.pl*** (Appendix A: BLAST_kegg.pl) .

It is worth noting that intraserovar analysis between *Salmonella* serovar Dublin strains was not performed because the reference annotation from Salmonella serovar Dublin strain CT_02021853 was not annotated in a consistent format throughout the genome, making the pathway assignment difficult. Arguably, the annotation could have been reformatted and used in the analysis but this would have been a time consuming process because the inconsistent nature of the annotation meant that manual reformatting would have been required.

3.2.1.2 Pathway mapping

Each functional gene was mapped to the corresponding gene from *Salmonella Typhimurium* LT2 using the ***reciprocal_fasta.pl*** script described in section 2.2.3 (Appendix A: Reciprocal_fasta.pl). This script also assigned KEGG pathways to each gene based on the assigned orthologs from LT2, the LT2 pathways are publicly available from KEGG as .list files. Any genes which did not hit LT2 were mapped against the reference genome for KEGG transferral.

Finally the genes which also didn't map to the reference were BLASTed against all KEGG sequences. If the hit was above 75% identity and 85% in length it was accepted as an ortholog. Blast_kegg.pl was then used to take the KEGG Gene ID from the hit, parse the corresponding genome .list to assign the correct pathways Appendix A: BLAST_kegg.pl).

KEGG doesn't store pathway data on pseudogenes, and the ***reciprocal_fasta.pl*** script uses amino acid sequences. Sequence comparisons for pseudogenes needed to be at the DNA level to allow for frameshifts. To assign KEGG data to pseudogenes a BLASTn was performed for every pseudogene against all KEGG sequences. If the hit was above 75% identity and 85% in length it was accepted as an ortholog and the pathways were mapped to the pseudogene using BLAST_kegg.pl as described above.

The pathway counts were converted into percentages for visualisation purposes, so that functional pathways could be easily compared to pseudogene pathways. These were sorted into descending order by functional pathways and visualised as radial diagrams.

3.2.1.3 Higher KEGG description mapping

In the KEGG database all pathways are sorted into hierarchical groups based on their functions (http://www.genome.jp/kegg/pathway.html). The pathways results were sorted into these groups to look at patterns in pathways loss from the first tier of the KEGG hierarchy. This was performed as 3.2.1.2 but with the higher level of KEGG description. Higher KEGG descriptions were also analysed because they can give an overview of general areas of gene loss which wouldn't be apparent when looking at a specific pathway.

3.2.1.4 Enrichment testing

Counts for all descriptions (KEGG first tier and pathways) were made using the ***test_BLAST_kegg.pl*** script (Appendix A: test_BLAST_kegg.pl) The R script ***updated_analysis.R*** then used the CORNA package [194] to sample the

hypergeometric distribution and perform fisher's exact test on the data (Appendix A: updated_analysis.R).

In each genome the combined totals for functional and non-functional pathways were used as the population. From this we could predict how many genes we would expect to see associated with a particular pathway by random chance. This can then be compared to the quantities we actually observe. A p-value is calculated as an indication of statistical significance, for multiple comparisons it is standard to adjust the p-value because the rate of false positives is higher thus an unadjusted p-value will give a much less stringent threshold [195]. For this particular analysis the results have been ordered by statistical significance (both the standard and adjusted p-values have been included), this is a good starting point for examining areas of greatest pathway loss/pathway preservation.

3.2.1.5 Proof of concept of the network analysis of the entire KEGG network

The entire KEGG network was downloaded from the KEGG website in KGML format. The KEGGtranslator tool was used to convert the network into OWL format. The output kept the most of the node information but many of the interactions were lost during conversion. KEGGtranslator was then used to convert the original network to sif format. Nodes which represented multiple enzymes only kept the annotation for one enzyme, but the interactions were intact. The sif file was opened in BioLayout, and clustering was performed to see how well the network clusters [196]. The clustering was performed using the inbuilt MCL (Markov Clustering) function. The network clustered into over three thousand clusters with many of the nodes remaining unassigned. This is due to some metabolites being very ubiquitous across the entire network (for example H_20 and ATP). BioLayout was used to identify these nodes by ordering all nodes by the number of interactions. Nodes that were very promiscuous (having a disproportionately high number of interactions) were pruned from the network to see if this improved the clustering. Opening the pruned network showed that some sub-networks had formed (between 5 and 15 nodes in size), these had no interaction with the main network any longer. The MCL function was used to cluster the pruned network. The clustering assigned most nodes

into a cluster this time but the cluster structure consisted of 5 or more metabolites and one or two gene nodes.

For the proof of concept a sub-network was used as the previous network was not clustering well. The sub-network was made from the Microbial metabolism in diverse environments overview network. This network was chosen because it was considerably smaller than the entire network, did not contain superfluous compounds and the KGML format node assignment was in a more consistent format. The network was pruned to only include gene-gene interactions, this was purely to reduce the network into a manageable size. A mock set of pseudogenes were then manually assigned to the network. The network was clustered a before using the MCL function. The clustering produced nine clusters with a minimum size of six genes in node. The cluster lists were extracted. The list, consisting of pseudogenes, functional genes and the cluster that they group into were analysed for enrichment using Fisher's exact test (as described in 3.2.1.4).

3.2.1.5.1 Mapping pseudogenes onto the network

As the proof of concept used a relatively small network the mapping of pseudogene could be performed manually, however, in life size data set this would not be possible. For the given time frame mapping the pseudogenes to the network was not feasible. The previous work in this chapter produced lists of genes and pseudogenes mapped to KO (kegg orthology) IDs. The sif format does not seem to have a consistent way of annotating enzymatic nodes, some have common gene names, E.C number or truncated protein product. The fact that each node only held information on one enzyme meant that a lot information would be lost.

In order to get all of the node information the KGML format would need to be parsed. This could be parsed into a hash where the KO ID would be the key and any KOs that share a node would be recorded as values. Files which have KO ID mapped to gene name and protein product could be downloaded from the KEGG server (subject to paying their subscription fees). These could then be integrated into the aforementioned hash. The hash could then be iterated through with the pseudogene list, if a pseudogene is present in any of the values for the hash then that ID is

recorded as having a pseudogene and added to a list. This list would be just the nodes that contain an enzyme with a pseudogene.

A script could then be written that would go through the sif file and the list of pseudogene nodes systematically looking for matches between the sif node and the pseudogene list. If the node does contain a pseudogene this mode would be assigned to the pseudogene class. The assignment of the pseudogene class means that a sub-network can be viewed.

3.2.2 Results

This section is divided into intraserovar comparisons, to ascertain the suitability of representative strains and interserovar comparisons to look at similarities and differences between different serovars. The full results for this section are available Appendix A: KEGG_summary.xlsx

3.2.2.1 Intraserovar comparisons

3.2.2.1.1 Gallinarum

Strain 287/91 had more pseudogenes than SG9, 309 and 276 respectively. There were 35 different pathways involving pseudogenes in 287/91 and 38 in SG9.

Table 6 shows that there were six unique pathways in Gallinarum SG9 all associated with one gene (SG9_2441). There were three unique pathways in 287/91 assigned to 3 separate genes.

The Gallinarum strains show high similarity at the KEGG higher description level (Figure 17). Looking more specifically at pathways, there are some deviations between pseudogene formation in KEGG pathways, showing slight variation of pathway frequency rather than complete omission of a particular pathway. Figure 18 is in descending frequency of pseudogenes to pathways in Gallinarum SG9, the pseudogene frequency in Gallinarum 287/91 shows the same descending pattern but with slight deviations.

Table 6 Pathways showing pseudogene formation unique to one strain in the Gallinarum serovar after comparison between SG9 and 287/91 for pseudogene formation in pathways.

Strain	Pathway	Pathway Description	Pseudogene	Function
SG9	00280	Valine, leucine and isoleucine degradation	SG9_2441	3-ketoacyl-CoA thiolase
SG9	00281	Geraniol degradation	SG9_2441	3-ketoacyl-CoA thiolase
SG9	00362	Benzoate degradation	SG9_2441	3-ketoacyl-CoA thiolase
SG9	00410	beta-Alanine metabolism	SG9_2441	3-ketoacyl-CoA thiolase
SG9	00592	alpha-Linolenic acid metabolism	SG9_2441	3-ketoacyl-CoA thiolase
SG9	00642	Ethylbenzene degradation	SG9_2441	3-ketoacyl-CoA thiolase
287/91	00633	Nitrotoluene degradation	SG0855	nitroreductase A; K10678 nitroreductase [EC:1.-.-.-]
287/91	00920	Sulfur metabolism	SG1654	protein MalY (EC:4.4.1.8); K14155 cystathione beta-lyase [EC:4.4.1.8]
287/91	00020	Citrate cycle (TCA cycle)	SG4145	class I fumarate hydratase (EC:4.2.1.2); K01676 fumarate hydratase, class I [EC:4.2.1.2]

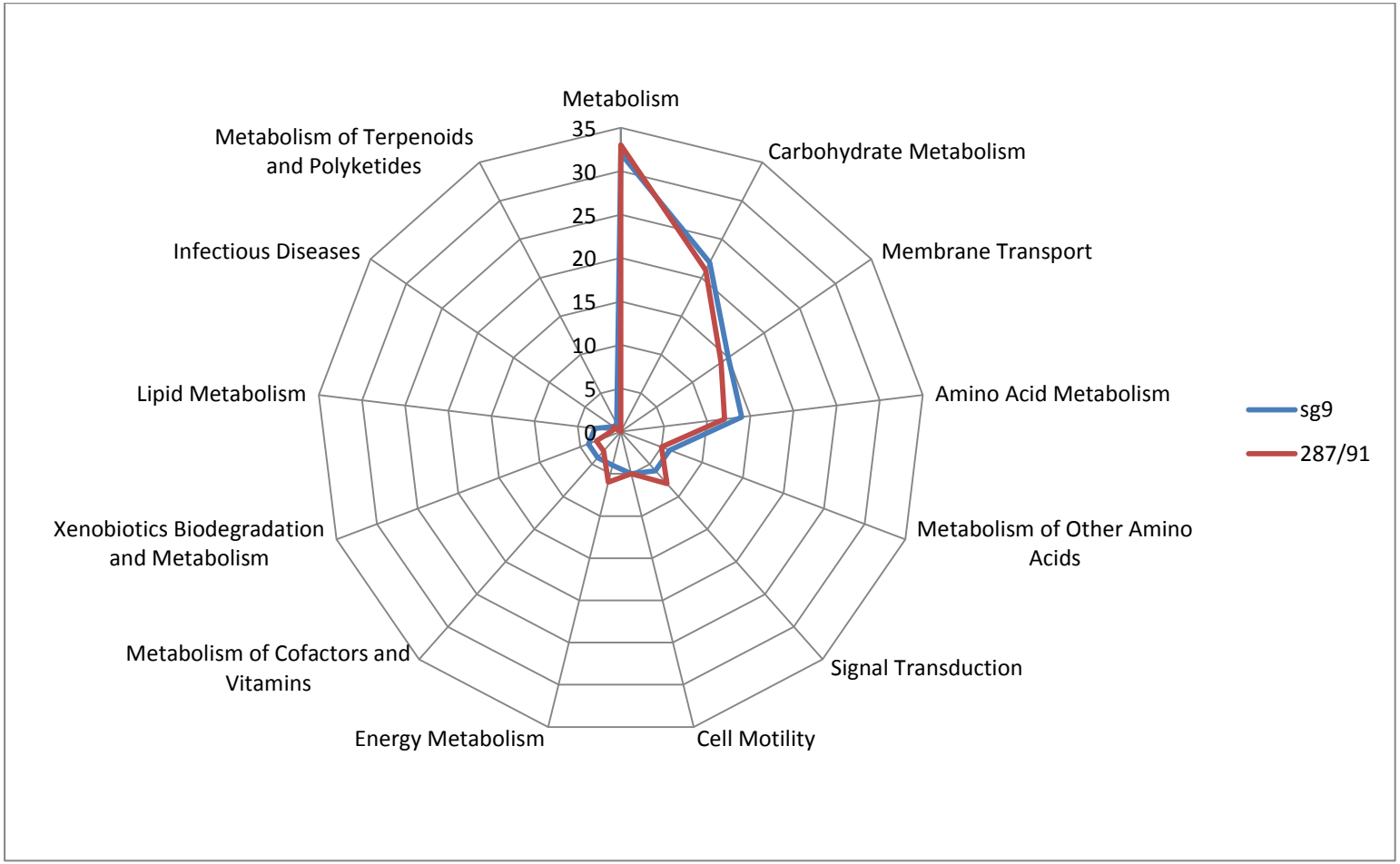


Figure 17 Comparison of higher KEGG pseudogene assignment between Gallinarum SG9 and 287/91

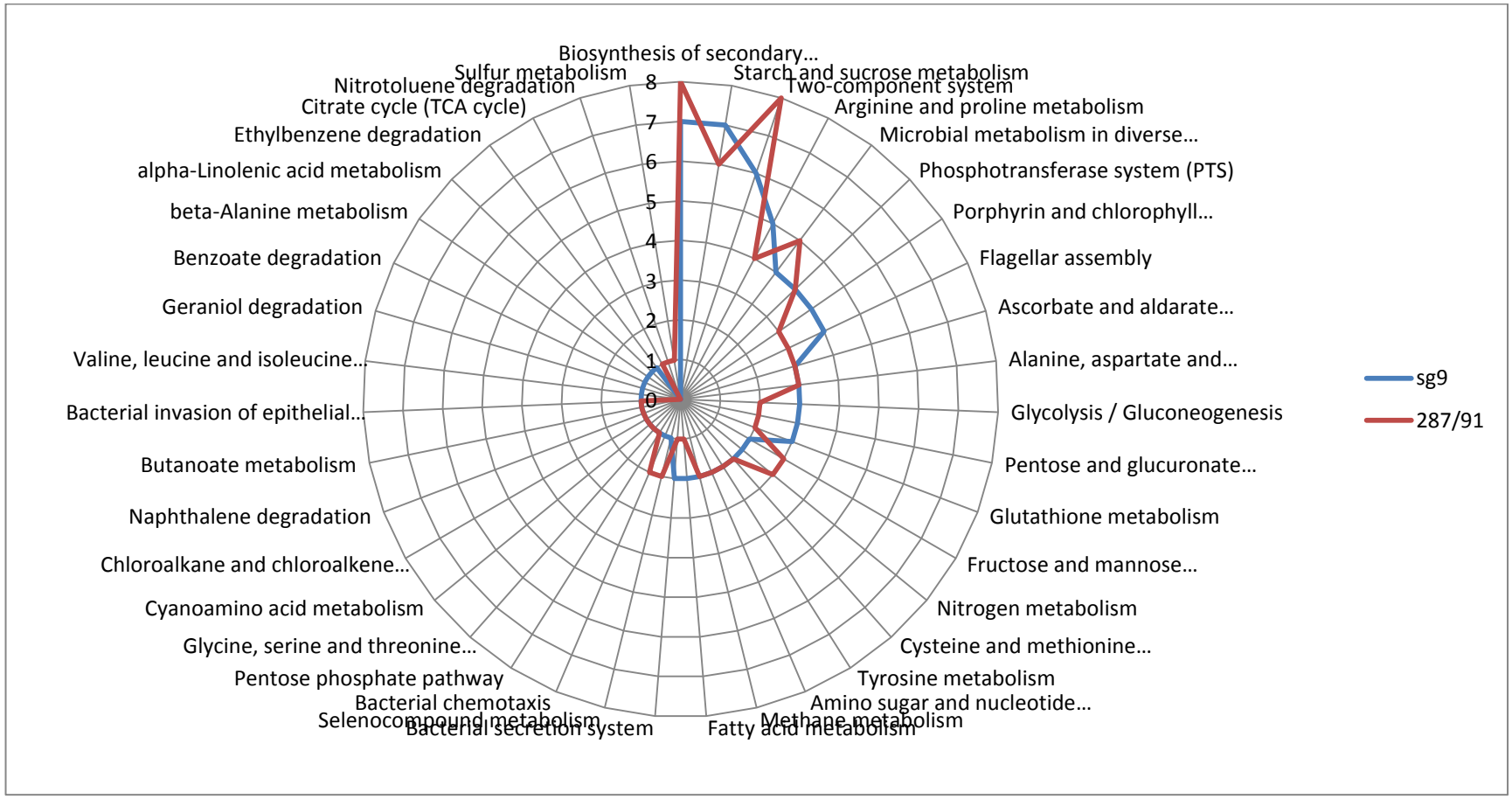


Figure 18 Comparison of frequency of pseudogenes in KEGG pathways between two Gallinarum strains, SG9 and 287/91, the order of pathways is based on decreasing frequency in SG9.

3.2.2.1.2 Choleraesuis

Choleraesuis strain SC-B67 had 154 pseudogenes compared to the 116 in Choleraesuis SCA50. There were no unique pseudogene pathways in Choleraesuis SCA50, Choleraesuis SC-B67 on the other hand, had 9 unique pseudogene pathways from 6 different genes.

The distribution of pseudogenes in higher KEGG terms showed a similar pattern but there were proportionally less for each description in Choleraesuis SCA50 (Figure 19). However, Figure 20 shows that the distribution between the two strains was variable at the KEGG pathway level.

Table 7 shows pathways with pseudogene formation unique to one serovar in the Choleraesuis strain.

Strain	Pathway	Pathway Description	Pseudogene	Function
SC-B67	00230	Purine metabolism	SCH_PS80	5'-nucleotidase/2',3'-cyclic phosphodiesterase
SC-B67	00230	Purine metabolism	SCH_PS83	bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase (EC:2.1.2.3 3.5.4.10)
SC-B67	00680	Methane metabolism*	SCH_PS37	esterase (EC:3.1.1.1)
SC-B67	00240	Pyrimidine metabolism	SCH_PS80	5'-nucleotidase/2',3'-cyclic phosphodiesterase
SC-B67	00620	Pyruvate metabolism	SCH_PS81	formate acetyltransferase 2; K00656 formate C-acetyltransferase [EC:2.3.1.54]
SC-B67	00640	Propanoate metabolism	SCH_PS81	formate acetyltransferase 2; K00656 formate C-acetyltransferase [EC:2.3.1.54]
SC-B67	00660	C5-Branched dibasic acid metabolism	SCH_PS72	acetolactate synthase 2 catalytic subunit
SC-B67	00670	One carbon pool by folate	SCH_PS83	bifunctional phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase (EC:2.1.2.3 3.5.4.10)
SC-B67	00760	Nicotinate and nicotinamide metabolism	SCH_PS80	5'-nucleotidase/2',3'-cyclic phosphodiesterase
SC-B67	00790	Folate biosynthesis	SCH_PS144	dihydropteroate synthase

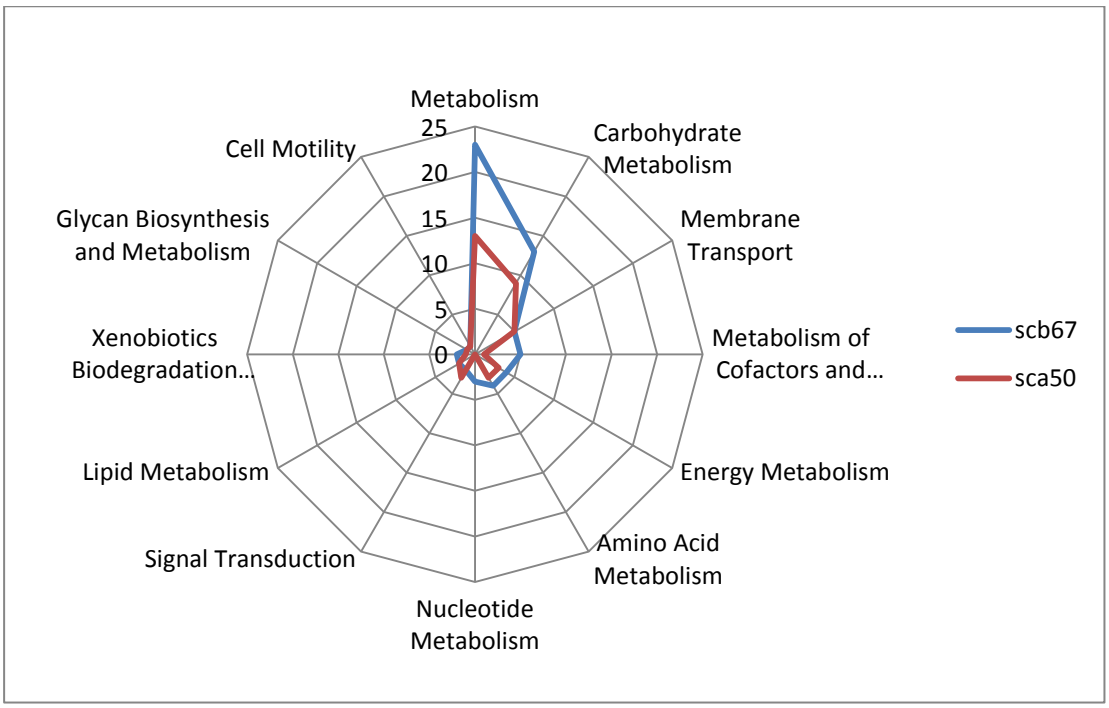


Figure 19 Comparison of higher level KEGG assignment of pseudogenes between two Choleraesuis strains (SC-B67 and SCA50) ordered by decreasing frequency in SC-B67.

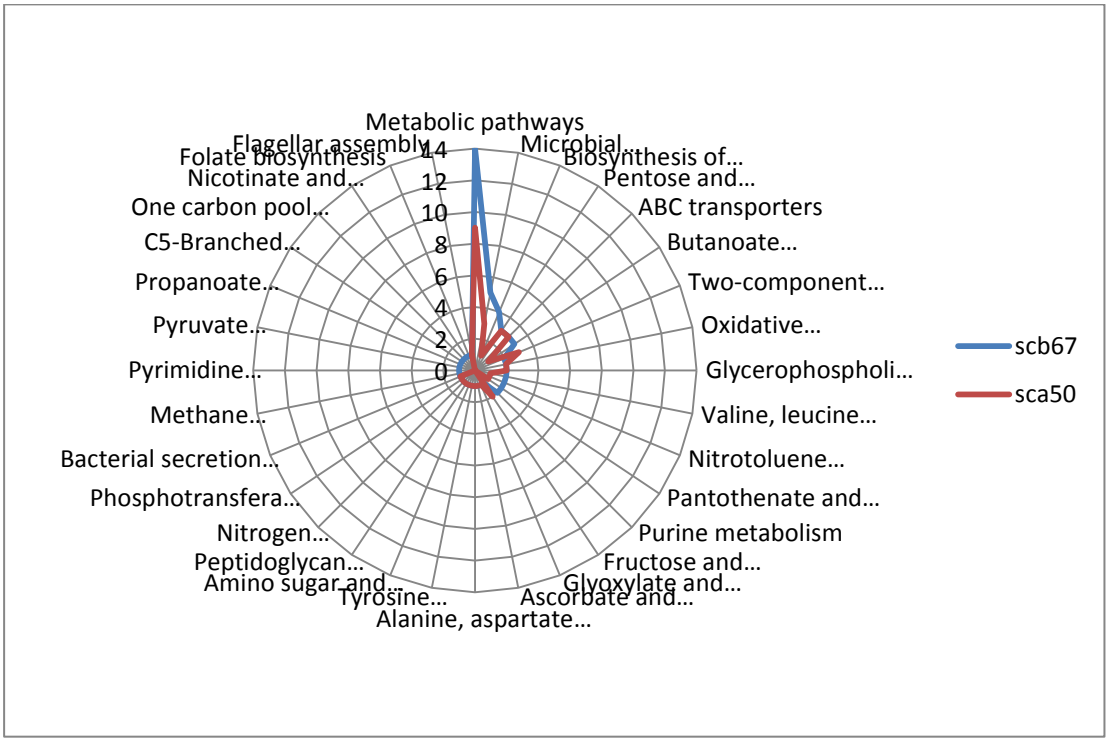


Figure 20 Comparison of pseudogene frequency in KEGG pathways between two different Choleraesuis strains (SC-B67 and SCA50) ordered by decreasing frequency in SC-B67.

3.2.2.2 Interserovar comparisons

The number of functional pathways vs. pseudogenes pathways was much bigger. With that in mind, the pathway counts were converted into percentages so that functional pathways could be easily compared to pseudogene pathways. These are presented on a logarithmic scale because some of the KEGG descriptions (such as 'Metabolism') were far more frequent, skewing the visualisation in a graph. These were ordered in descending order by functional pathways and visualised as radial diagrams. Figure 21 shows that all four serovars have a very similar distribution of pathways assigned to functional genes. The radial diagram in Figure 22 shows the distribution of pathways assigned to pseudogenes. It is in the same order as Figure 21, and shares the spiral pattern in the first four pathways. After this point the pathway distribution is unique across serovars and between functional and non-functional gene assignment within serovars.

Of 397 pathways in KEGG only 72 actually had genes assigned to them. This is probably because many of the pathways are Eukaryote or Mammal specific, such as 'Lysosome' and 'Taste Transduction' respectively.

For the 'higher pathways' rather than 397 there were only 50 more general terms which encompass the original terms. Of these 50 terms only 20 had genes assigned to them.

There were 53 different pathways (in both tiers) that have functional genes but no pseudogenes across all serovars (Appendix C). Many of the pathways were associated with genetic information processing, higher terms included 'Nucleotide Metabolism' and 'Replication and Repair', the KEGG pathways included 'Mismatch Repair', 'Protein Export' and 'DNA Replication'. The five most common terms (based on functional gene frequency) that had no pseudogenes were 'Nucleotide Metabolism', 'Purine Metabolism', 'Replication and Repair', 'Pyrimidine Metabolism' and 'Folding, Sorting and Degradation'. Other notable pathways such as 'Streptomycin Biosynthesis', 'Lipopolysaccharide Biosynthesis' and 'Novobiocin Biosynthesis' also showed no pseudogene formation

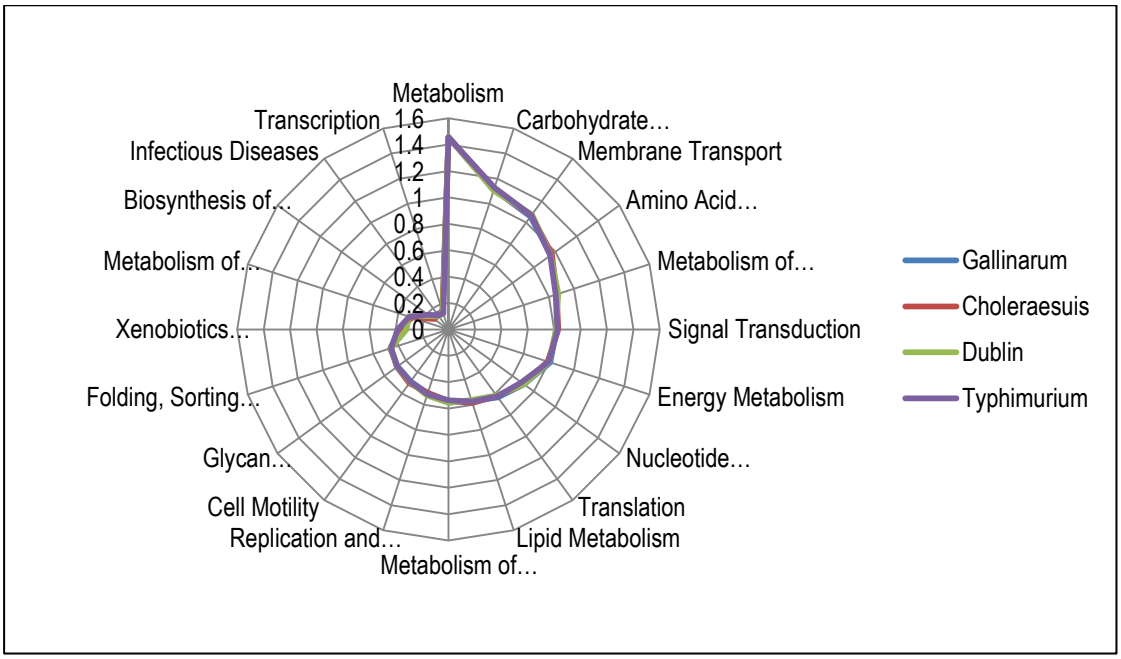


Figure 21 Percentages of functional genes in higher KEGG descriptions for each serovar

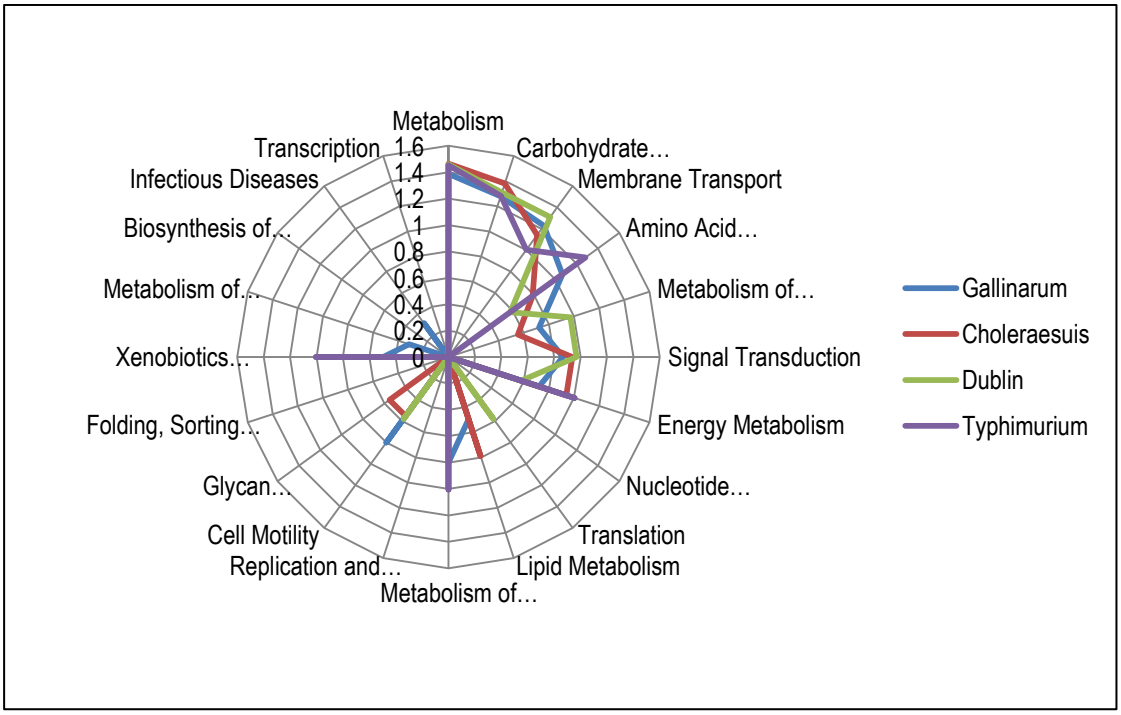


Figure 22 Percentages of pseudogenes in higher KEGG description for each serovar, in the same order as Figure 21.

11 pathways were found to have pseudogenes across all serovars, 9 of these were ‘Metabolism’ pathways. The other two were the ‘Phosphotransferase system (PTS)’ pathway which is part of the ‘Membrane Transport’ higher description.

3.2.2.2.1 Choleraesuis

The five pseudogene pathways unique to Choleraesuis SCA50 were ‘Oxidative phosphorylation’, ‘Glycerophospholipid metabolism’, ‘Pantothenate’ and ‘CoA biosynthesis’ and ‘Peptidoglycan biosynthesis’ which is part of the higher pathway ‘Glycan Biosynthesis and metabolism’ (Table 8).

There was one higher description and one pathway with pseudogene formation that were present across all serovars except for Choleraesuis; the higher description being ‘Metabolism of Other Amino Acids’, which contains the aforementioned pathway ‘Selenocompound metabolism’ (Table 9).

The pathways for Choleraesuis SCA50, ordered by significance are shown in Figure 23. None of the results showed any significance for the adjusted p-values. ‘Pentose and Glucronate interconversions’ was significant for the unadjusted p-value.

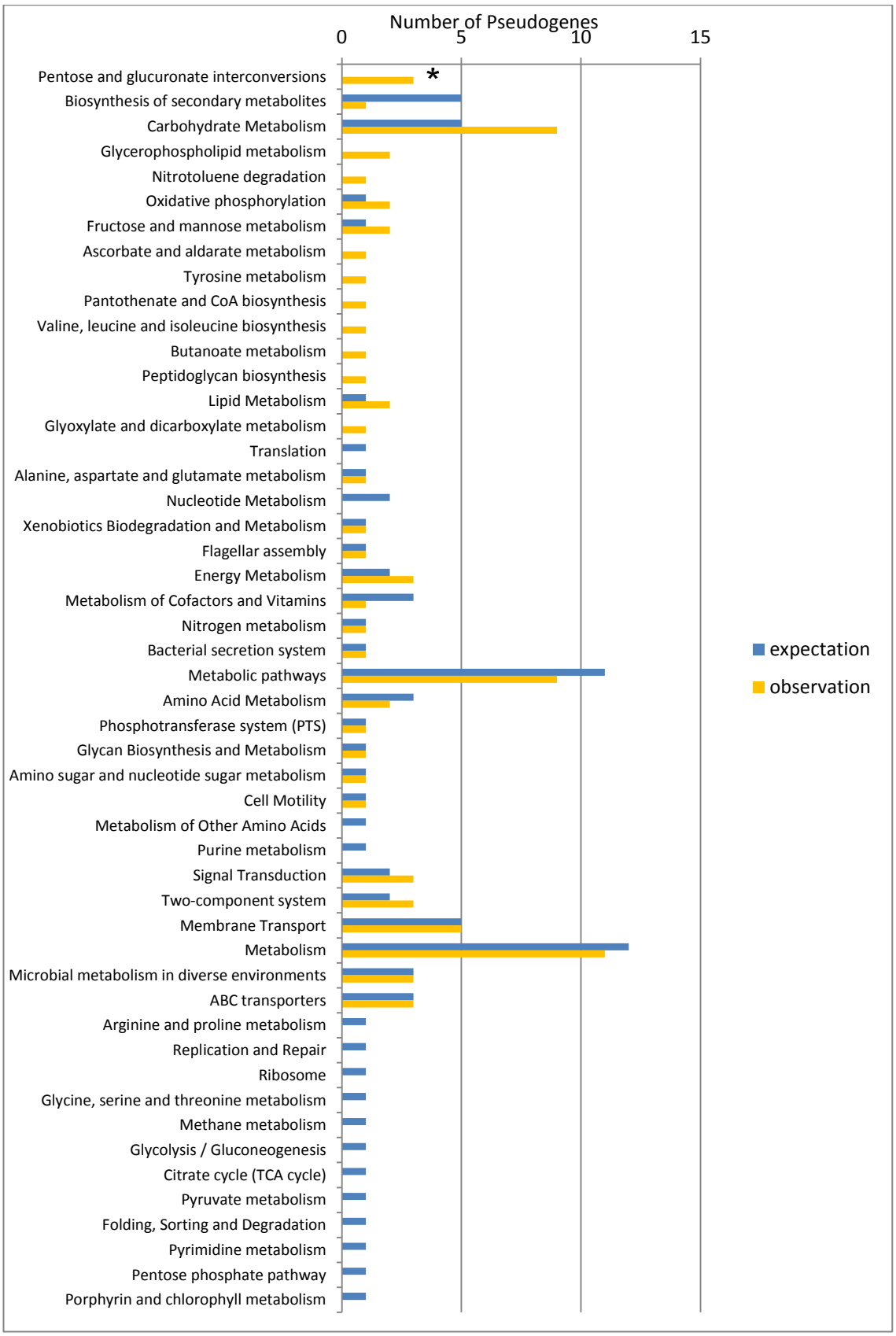


Figure 23 Choleraesuis SCA50 - Observed and expected pseudogene counts in KEGG pathways and higher descriptions calculated using Fisher's exact test, ordered by ascending p-value. Pathways with a '*' indicate significance before adjustment.

Table 8 Shows pathways which display Choleraesuis SCA50 specific pseudogene formation. Pink cells show less observed pseudogenes than expected (according to Fisher's exact test), Amber cells show identical observed and expected counts, Green cells show more observed pseudogene formation compared to that predicted by Fisher's exact test.

Pathway	Description	Functional Count				Expected Pseudo Count				Observed Pseudo Count			
		Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh
15	Glycan Biosynthesis and Metabolism	49	50	50	52	2	1	1	0	0	1	0	0
00190	Oxidative phosphorylation	41	41	42	41	2	1	1	0	0	2	0	0
00564	Glycerophospholipid metabolism	26	27	25	27	1	0	0	0	0	2	0	0
00550	Peptidoglycan biosynthesis	23	24	24	24	1	0	0	0	0	1	0	0
00770	Pantothenate and CoA biosynthesis	21	21	21	22	1	0	0	0	0	1	0	0

Table 9 Shows pathways only disrupted in Gallinarum Dublin and Typhimurium. Pink cells show less observed pseudogenes than expected (according to Fisher's exact test), Amber cells show identical observed and expected counts, green cells show more observed pseudogene formation compared to that predicted by Fisher's exact test.

Pathway	Description	Functional Count				Expected Pseudo Count				Observed Pseudo Count			
		Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh
9	Metabolism of Other Amino Acids	66	62	65	63	3	1	1	0	5	0	1	1
00450	Selenocompound metabolism	16	15	16	16	1	0	0	0	1	0	1	1

3.2.2.2.2 Dublin

Salmonella Dublin SD3246 had eight pathways and one higher description which only showed pseudogene formation in this serovar. The only higher description, Translation, holds the Ribosome pathway which is one of the eight pathways. The other 7 pathways were either 'Metabolism' pathways or 'Xenobiotics and degradation' pathways (Table 10).

There were no KEGG descriptions/pathways showing pseudogene formation that were present across all serovars except for Dublin.

There were no significant pathways according to adjusted Fisher's exact test. Those significant prior to adjustment were 'Toluene degradation', 'Chlorocyclohexane and chlorobenzene degradation' and 'Fluorobenzoate degradation'.

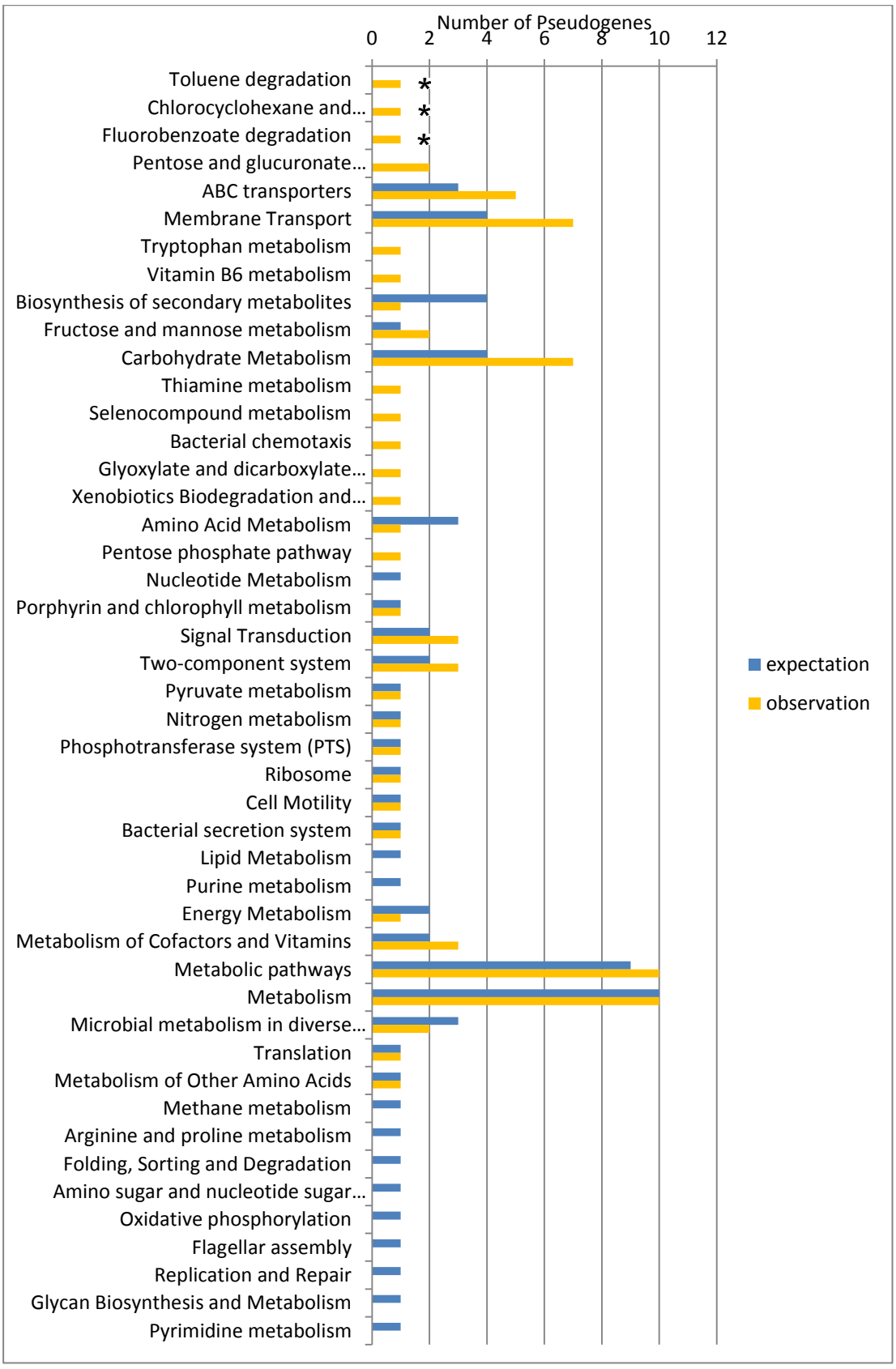


Figure 24 Dublin SD3246 - Observed and expected pseudogene counts in KEGG pathways and higher descriptions calculated using Fisher's exact test, ordered by ascending p-value. Pathways with a '*' indicate significance before adjustment.

Table 10 Shows pathways which display Dublin specific pseudogene formation. Pink cells show less observed pseudogenes than expected (according to Fisher's exact test), Amber cells show identical observed and expected counts, Green cells show more observed pseudogene formation compared to that predicted by Fisher's exact test.

Pathway	Description	Functional Count				Expected Pseudo Count				Observed Pseudo Count			
		Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh
6	Translation	82	79	78	84	3	1	1	0	0	0	1	0
03010	Ribosome	55	52	50	56	2	1	1	0	0	0	1	0
00620	Pyruvate metabolism	41	42	41	43	2	1	1	0	0	0	1	0
00730	Thiamine metabolism	12	12	13	13	1	0	0	0	0	0	1	0
00750	Vitamin B6 metabolism	10	10	10	10	0	0	0	0	0	0	1	0
00380	Tryptophan metabolism	9	9	9	9	0	0	0	0	0	0	1	0
00623	Toluene degradation	9	8	1	9	0	0	0	0	0	0	1	0
00361	Chlorocyclohexane and chlorobenzene degradation	2	2	1	2	0	0	0	0	0	0	1	0
00364	Fluorobenzoate degradation	1	1	1	1	0	0	0	0	0	0	1	0

3.2.2.2.3 Gallinarum

Gallinarum SG9 was the only serovar with a pathway, 'Starch and Sucrose Metabolism', that shows significant pseudogene formation after adjusting the p-value from the Fisher's exact test. There were 6 pathways that were flagged as significant prior to adjusting the p-values, they were 'Arginine and Proline Metabolism', 'Nucleotide Metabolism', 'Glutathione Metabolism', 'Ascorbate and Aldarate Metabolism', 'Ethylbenzene Degradation' and 'Flagellar Assembly'. It is worth noting that Nucleotide is under represented, that is there were less pseudogenes in this pathway than expected according to Fisher's exact test.

There were 16 pathways and two higher descriptions that showed pseudogene formation in Gallinarum SG9 only (Table 11). The two higher pathways were 'Infectious Diseases' and 'Metabolism of Terpenoids and Polyketides' both only had one pseudogene. With that in mind the pathways they encompassed were 'Bacterial Invasion of Epithelial Cells' and 'Geraniol Degradation'; respectively making those Gallinarum specific in terms of pseudogene formation too. The other 14 pathways could be broadly grouped into, 'Amino acid metabolism' (6 pathways), 'Xenobiotics Biodegradation and Metabolism' (4 pathways), 'Lipid Metabolism' (two), 'Carbohydrate Metabolism' ('Glycolysis/Gluconeogenesis') and, 'Energy Metabolism' ('Methane metabolism').

There were no KEGG descriptions/pathways showing pseudogene formation that were present across all serovars except for Gallinarum.

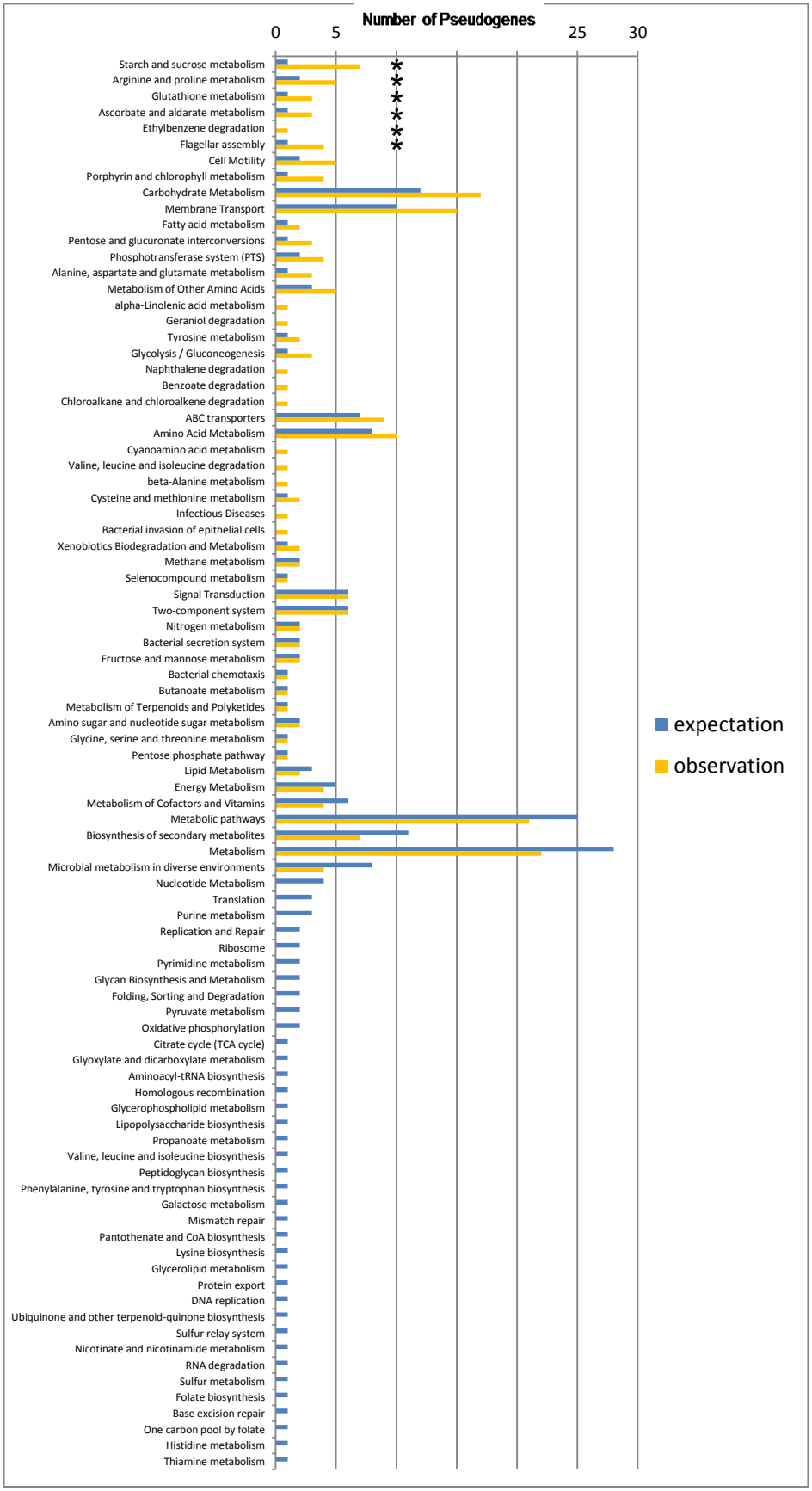


Figure 25 Gallinarum SG9 - Observed and expected pseudogene counts in KEGG pathways and higher descriptions calculated by Fisher's exact test, ordered by ascending p-value. Pathways with a '*' indicate significance before adjustment.

Table 11 Shows pathways which have Gallinarum specific pseudogene formation. Pink cells show less observed pseudogenes than expected (according to Fisher's exact test), Amber cells show identical observed and expected counts, Green cells show more observed pseudogene formation compared to that predicted by Fisher's exact test.

Pathway	Description	Functional Count				Expected Pseudo Count				Observed Pseudo Count			
		Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh
00330	Arginine and proline metabolism	43	44	41	44	2	1	1	0	5	0	0	0
00680	Methane metabolism	37	31	34	38	2	1	1	0	2	0	0	0
00010	Glycolysis / Gluconeogenesis	35	34	33	37	1	1	0	0	3	0	0	0
00480	Glutathione metabolism	18	18	19	18	1	0	0	0	3	0	0	0
00071	Fatty acid metabolism	12	13	11	13	1	0	0	0	2	0	0	0
00260	Glycine, serine and threonine metabolism	31	31	32	31	1	1	0	0	1	0	0	0
19	Metabolism of Terpenoids and Polyketides	26	24	27	27	1	0	0	0	1	0	0	0
05100	Bacterial invasion of epithelial cells	10	9	8	9	0	0	0	0	1	0	0	0
14	Infectious Diseases	10	9	8	9	0	0	0	0	1	0	0	0
00280	Valine, leucine and isoleucine degradation	9	9	8	10	0	0	0	0	1	0	0	0
00410	beta-Alanine metabolism	9	9	8	9	0	0	0	0	1	0	0	0
00460	Cyanoamino acid metabolism	9	8	9	8	0	0	0	0	1	0	0	0
00362	Benzoate degradation	6	6	6	6	0	0	0	0	1	0	0	0

00625	Chloroalkane and chloroalkene degradation	6	5	5	6	0	0	0	0	1	0	0	0
00626	Naphthalene degradation	5	5	4	5	0	0	0	0	1	0	0	0
00281	Geraniol degradation	4	4	4	4	0	0	0	0	1	0	0	0
00592	alpha-Linolenic acid metabolism	3	3	3	3	0	0	0	0	1	0	0	0
00642	Ethylbenzene degradation	1	0	0	0	0	0	0	0	1	0	0	0

3.2.2.2.4 Typhimurium

There were two pathways with pseudogene formation that were unique to *Salmonella* Typhimurium 4/74, namely, 'Sulfur Metabolism' and 'C5-Branched acid metabolism' (Table 12).

The data was also interrogated for pseudogene formation across all serovars bar Typhimurium. Eight pathways were found, these could be indicative of Typhimurium's method of infection (Table 13). The pathways included 'Fructose and mannose metabolism', 'Two-component system' and 'Cell Motility'.

Using the unadjusted p-values there were four significant pathways for pseudogene formation according to the unadjusted Fisher's exact test. Namely, 'Nitroluene degradation', 'C5-Branched dibasic acid metabolism', 'Sulfur metabolism' and 'Selenocompound metabolism'. None of these were significant after adjusting the p-values.

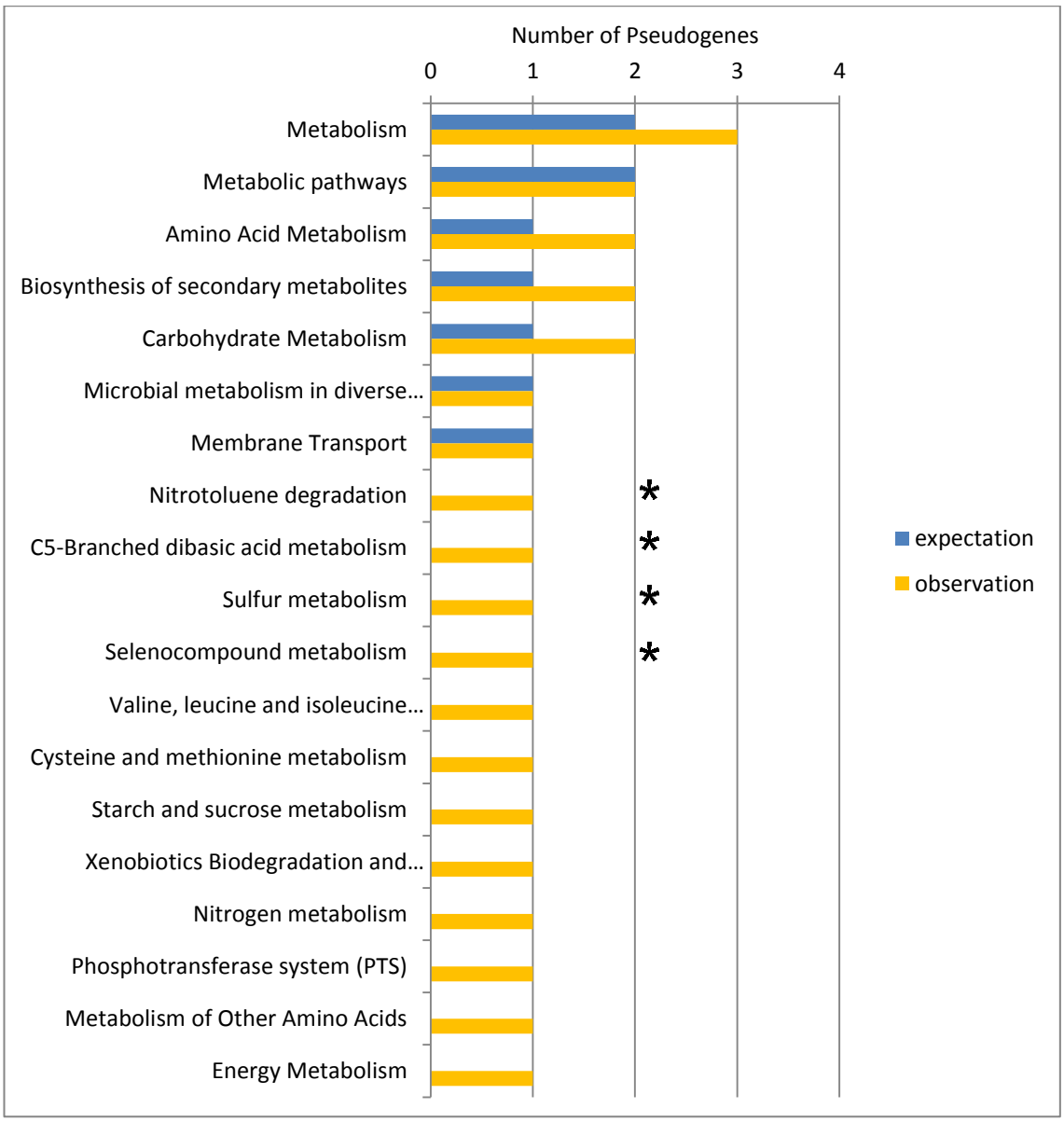


Figure 26 Typhimurium ST4/74 - Observed and expected pseudogene counts in KEGG pathways and higher descriptions calculated using Fisher's exact test, ordered by descending observed count.

Table 12 Pseudogene pathways only found in Typhimurium 4/74. Pink cells show less observed pseudogenes than expected (according to Fisher's exact test), Amber cells show identical observed and expected counts, Green cells show more observed pseudogene formation compared to that predicted by Fisher's exact test.

Pathway	Description	Functional Count				Expected Pseudo Count				Observed Pseudo Count			
		Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh
00920	Sulfur metabolism	14	14	14	15	1	0	0	0	0	0	0	1
00660	C5-Branched dibasic acid metabolism	12	11	12	12	0	0	0	0	0	0	0	1

Table 13 Shows host specific pathway loss, that is pathways only lost in Gallinarum, Choleraesuis and Dublin. Pink cells show less observed pseudogenes than expected (according to Fisher's exact test), Amber cells show identical observed and expected counts, Green cells show more observed pseudogene formation compared to that predicted by Fisher's exact test.

Pathway	Description	Functional Count				Expected Pseudo Count				Observed Pseudo Count			
		Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh
02010	ABC transporters	165	171	168	175	7	3	3	0	9	3	5	0
02020	Two-component system	137	142	136	147	6	2	2	0	6	3	3	0
5	Signal Transduction	137	142	136	147	6	2	2	0	6	3	3	0
7	Metabolism of Cofactors and Vitamins	154	153	160	159	6	3	2	0	4	1	3	0
16	Cell Motility	51	54	52	54	2	1	1	0	5	1	1	0
00051	Fructose and mannose metabolism	46	47	47	55	2	1	1	0	2	2	2	0
03070	Bacterial secretion system	44	42	54	45	2	1	1	0	2	1	1	0
00040	Pentose and glucuronate interconversions	27	27	27	29	1	0	0	0	3	3	2	0

Table 14 Shows pathways which have no pseudogenes across all hosts. Pink cells show less observed pseudogenes than expected (according to Fisher's exact test), Amber cells show identical observed and expected counts, Green cells show more observed pseudogene formation compared to that predicted by Fisher's exact test.

Pathway	Description	Functional Count				Expected Pseudo Count				Observed Pseudo Count			
		Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh	Gall	Chol	Dub	Typh
10	Nucleotide Metabolism	104	101	102	99	4	2	1	0	0	0	0	0
00230	Purine metabolism	80	77	78	75	3	1	1	0	0	0	0	0
17	Replication and Repair	57	55	59	59	2	1	1	0	0	0	0	0
00240	Pyrimidine metabolism	53	51	51	51	2	1	1	0	0	0	0	0
3	Folding, Sorting and Degradation	49	47	47	48	2	1	1	0	0	0	0	0

3.2.3 Discussion

The comparison between observed and predicted pseudogene frequency can indicate non-essential and essential pathways for host adaptation. With the genomes published in chapter 2, we have a unique opportunity to compare host adapted serovars with a host generalist.

The null hypothesis is that mutations within a genome can be considered as a randomly occurring phenomenon; therefore, if there were no essential genes in the genomes we would expect to see a pattern of mutation formation similar to the frequency of functional genes. That is if there are a lot genes linked to carbohydrate metabolism, for example, then we would expect to see gene loss for these pathways to be high based purely on chance mutation. If however there is some kind of selection pressure we would expect to see a much lower frequency of pseudogene pathways for essential genes. Figure 21 and Figure 22 show that pseudogene formation is not completely random and that there is a selection pressure on some pathways. For example, higher description 'nucleotide metabolism' and pathway 'replication and repair' have no pseudogenes. These pathways are involved in core functions rather than pathogenesis, the data supports that any gene loss in these pathways would be detrimental.

The suitability of using KEGG orthology and pathways is questionable, this analysis has shown inconsistencies in their pathway assignment quality and the fact that the KEGG ftp site is now only accessible for a fee is disappointing.

It is worth noting that pseudogene formation does not confer complete loss of functionality, there could be alternate routes in a given pathway or even paralogs in the genome, with that in mind the next step to these analyses would be to explore the pathways of interest in more detail to actually look for paralogs and alternative routes (Figure 27 and Figure 28).

In terms of in silico methods stochiometric formulas for known reactions can be integrated with metabolic pathways and flux values. Incorporating this into a

mathematical model can be used to predict the optimum route for a given pathway [Beste 2010].

Not only can these models predict the potential resilience of a given pathway to pseudogene formation but they actually used the technique to predict errors in the annotation based on information from pathways, It has been shown that in the glycerol utilization pathway is dependent on the glpK gene. This is contrary to the genome annotation which showed several alternative routes. Knock-out of the glpK gene confirmed that the model was correct, and glpK is essential for growth on glycerol [197]. Beste et al. were able to prove that the other genes are misannotated as alternate routes [198, 199].

Further to this tools such as Acorn, are available now for integrating simulations of genome scale metabolic reaction networks with metabolic pathways and mapping them to genome annotation and from that predicting the essentiality of each gene [200].

Another possible method of large scale analysis would be the use of phenotypic arrays such as Biolog. Biolog looks at the cell growth of a given strain in different substrates. This has been used with Salmonella strains to examine functional reduction of pathways across different strains [201]. One could envision an experiment where genes are systematically knocked out for a given pathway. Each mutant could then be tested to see how it grows on different media. From this it would be possible to see how well the mutant grows in substrates associated with the pathway of interest.

3.2.3.1 The suitability of representative serovars

The intraserovar comparison is of interest because the previously sequenced strains do not have well defined virulence. They were isolated from infected livestock some years ago and have since been growing in a lab. Conversely, our serovars have known virulence, so finding areas of difference/similarity will shed light on the suitability of the current reference strains as representatives of their serovar.

The fact that the pseudogene formation was similar between strains in the host-specifist serovar (Gallinarum) and showed deviance in the host-restricted (Choleraesuis) serovar could be explained by their host specificity. As serovars become more host specific we see a higher level of gene attrition. As Choleraesuis can infect more than one different host, perhaps it is in the process of becoming host specific and shows an intermediate level of gene loss, where different pseudogenes have formed in different strains. For example, Choleraesuis SC-B67 has more pseudogenes, including shdA which is a well-known host colonisation factor. Mutation studies have shown that it is not needed in pigs [202], yet it is still a fully formed gene in Choleraesuis SCA50 (SCA50_2684) but has a frame shift in Choleraesuis SC-B67 (SCPS112). The fact that the gene is still functional in Choleraesuis SCA50 is not indicative of its requirement in the genome. The pseudogene formation in SC-B67 but lack of attrition in SCA50 could be indicative that SC-B67 has been host-adapted for longer.

Another limitation with this comparison is that it is based on publically available annotations from different labs. This will contribute to some of the discrepancies. For example six of the pseudogenes in Choleraesuis SC-B67 were annotated as selenocysteine read-throughs in Choleraesuis SCA50's orthologs. In terms of DNA these gene sequences are identical between strains, it is a question of which annotation one accepts when performing the comparisons. The only sure way of getting consistency for this type of comparison would be to annotate both genomes by exactly the same method.

Host specific servoars seem to be streamlined enough that they predominantly carry genes that are specific to a small niche meaning that in theory any strain from that serovar could be used as a representative. The general idea of defining representative strains for a serovar is questionable, especially when the serovar has a broader host specificity or is in the transition towards host specificity. I believe that advancements could be made through combining experimental data such as transcriptomics and mutagenesis to find a core gene set for a given serovar's host phenotype. For example in *E.coli* a systems biology approach has been used to investigate its

tolerance to acid. This integrated mutation studies with metabolic pathways, transcriptomics and phenotypic analysis [203].

As discussed, there were several limitations to this exercise, firstly the comparisons only consisted of two genomes from the serovar. It is recognised that some serovars are atypical of their expected phenotype [23]. Repeating this process now may be more informative as there are more Choleraesuis and Gallinarum genomes available in the public domain. Secondly, the annotation and assembly of our four serovars took considerable time. As there are now more tools available for assembly and annotation and bacterial sequencing is considerably cheaper it is more feasible to actually perform a more extensive analysis.

The premise of serovars and genovars is in some respects limited. These definitions are manmade groupings, evolution doesn't fall neatly into groups. It is complex to fathom how many serovars would be needed for a comprehensive analysis of host specificity. Examining the pan-genome of *Salmonella* could offer some clues. By plotting all of the available genomes against the cumulative count of genes families one can look at where the graph plateaus, this gives an indication of how many genomes are required to see most of the genes representative of *Salmonella.* Recently, the *Salmonella* pan genome was described as a 'closed genome', meaning that adding a new serovar to the pan genome plot shows a jump in novel gene families, however, adding a second isolate for the same serovar has little effect on the number of gene families [20]. Other work with Salmonella has shown that within a given serovar there is a tendency towards low diversity [48]. For example, Holt *et al.* used 17 strains of *S.typhi* for SNP analysis, they described typhi as having low recombination and needed a higher number of strains in order to obtain SNP data [22]. When looking at the genes rather than polymorphisms far fewer than 17 strains would have been needed. From these experiments we can infer that although only four serovars seems like a low number of representatives perhaps adding more wouldn't enhance the study in terms of increasing our knowledge of host specificity. Nevertheless, adding more isolates of the same serovar would definitely increase the confidence of the results.

Another method of adding more gravitas to the intraserovar comparison would be to define the virulence of the reference strains Choleraesuis SC-B67 and Gallinarum 287/91. This could be achieved by trying to infect pigs and chickens and observing the effects of infection.

3.2.3.2 Universal preservation of genetic processing genes and loss in metabolism

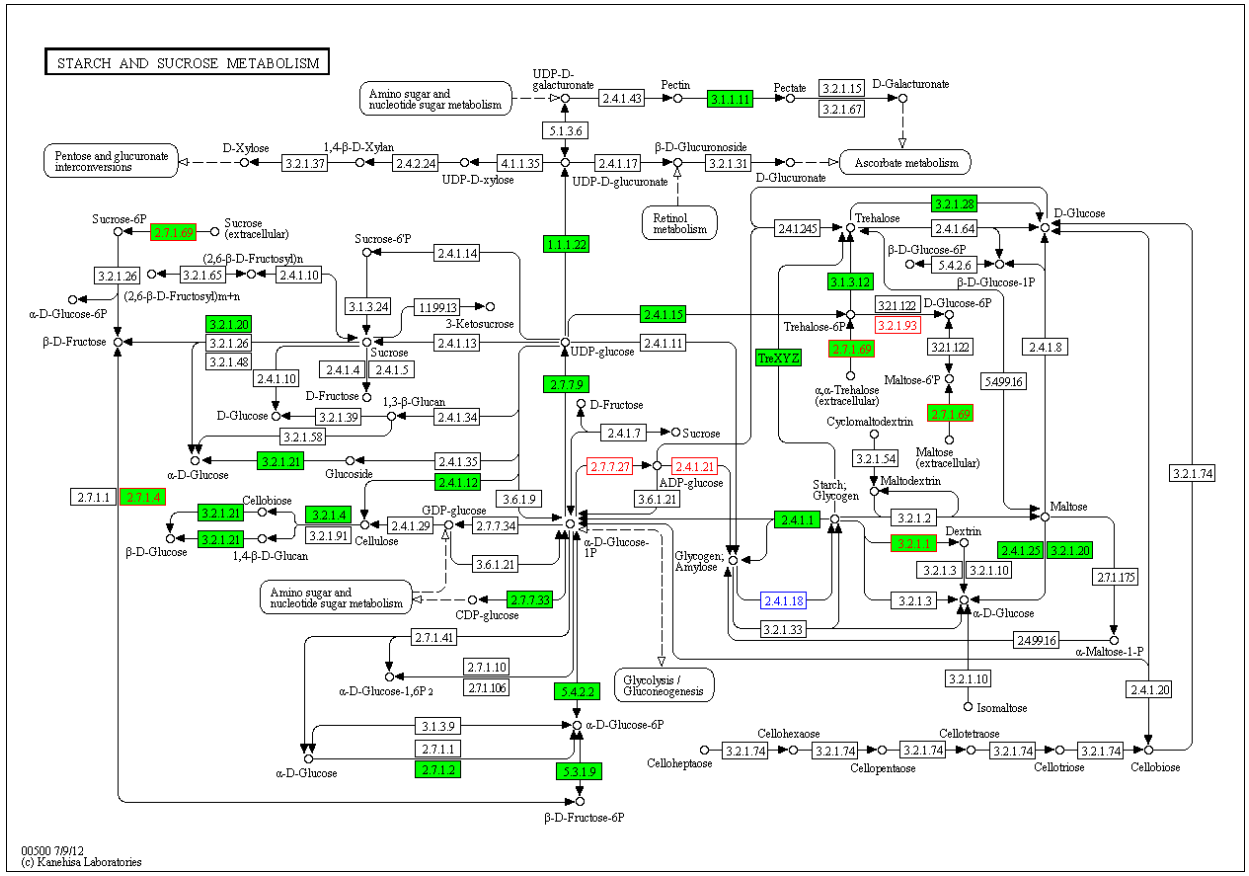
The comparison of pseudogene formation across all our serovars shows that it is variable, however there is a trend towards the preservation of certain pathways. The radial diagram (Figure 22) shows that there is no pseudogene formation in 'nucleotide metabolism', 'replication and repair' and 'folding sorting and degradation'. These are all involved in genetic information processing, these would all fall into the core genome and would likely be highly detrimental to the organism if disrupted.

Converse to this most of the variation between pseudogenes is in different types of metabolism, this concurs with the paradigm that *Salmonella* serovars are extremely flexible in terms of nutrient use, previous studies have shown that they have a lot of scope for enzyme loss [204, 205].

3.2.3.3 Role of Starch and Sucrose Metabolism in Gallinarum

This pathway shows significant enrichment in gene loss in Gallinarum SG9 and checking this against the Gallinarum 287/91 strain shows identical pseudogene/absence in this pathway. The pathway map shows the regions of gene loss (Figure 27). Gene loss/absence in the glg(ABC) genes (some of the genes lost in this pathway) have been recorded before in other chicken specific serovars [26]. Further to this there have been links between glgC and extra-host survival, that is survival outside of the host as these sugars are packed into endogenous stores for times of starvation [206]. This could concur with the way that Gallinarum is spread, compared to host generalists that cause diarrhoea and have high levels of shedding.

To test this hypothesis a combination of gene mutation/recombinase would be required in Typhimurium and Gallinarum respectively. If the mutated Typhimurium strains show reduced survival in the extrahost environment or the Gallinarum strains show increased survival this would be indicative of the role of starch and sucrose metabolism in survival outside of the host.



LEGEND

1.1.1.1 Present in LT2 and SG9

1.1.1.1 Present in LT2, all orthologs are pseudogenes in SG9

1.1.1.1 Present in LT2, no orthologs in SG9

1.1.1.1 Present in LT2, functional and pseudogene orthologs in SG9

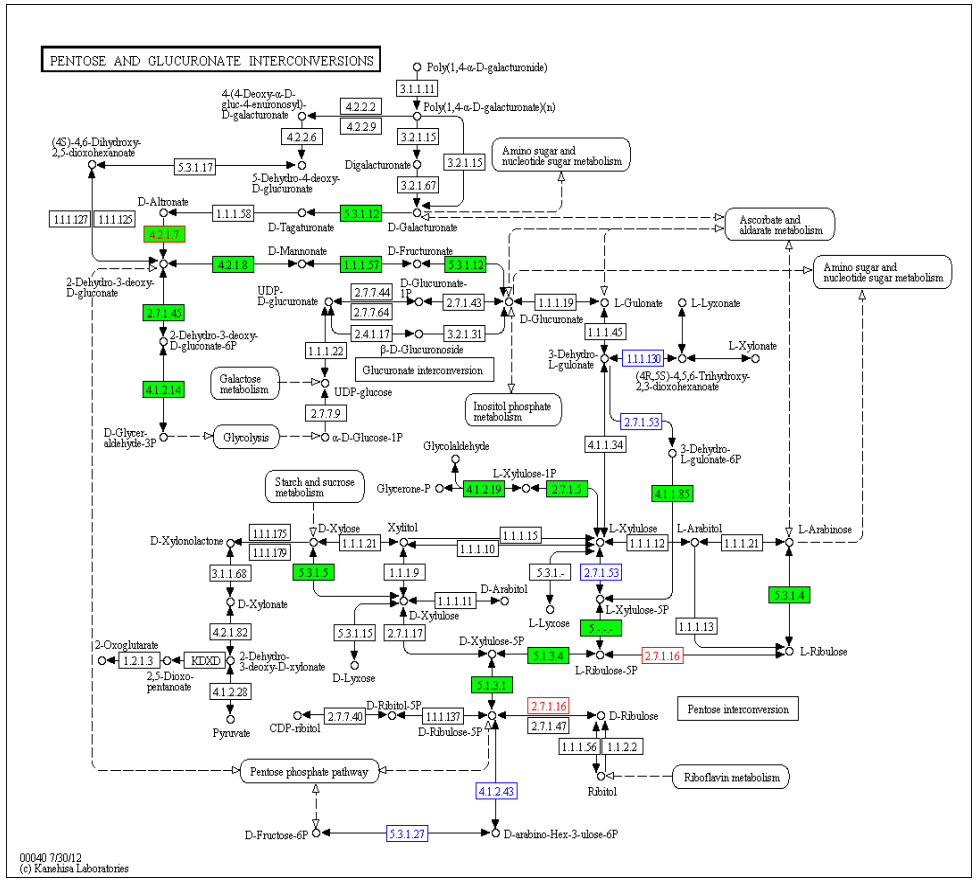
Figure 27 KEGG pathway map of Starch and Sucrose Metabolism. Green shows functional genes common to Gallinarum SG9 and Typhimurium LT2. Red outline shows pseudogene formation in Gallinarum SG9 and Blue outline shows gene absence in SG9. Note that some green areas with a red outline, this occurs when this there is a version of a functional gene and a pseudogene (presumably paralogs).

3.2.3.4 Pentose and Gluconorate interconversion in Choleraesuis/pigs

Both typhimurium and choleraesuis infect pigs, however it has been shown that typhimurium grows very quickly, causes an immune response, and is cleared by the host. In contrast, choleraesuis grows much more slowly and goes on to cause a systemic infection [137]. Counter-intuitively, this may be due to choleraesuis being less well adapted to the environment of the pig gut, enabling it to grow more slowly and therefore avoid the immune response that clears typhimurium.

Pentose and Gluconorate showed pseudogene enrichment in Choleraesuis SCA50 (Figure 23), closer inspection of this pathway shows that there is a lot of gene absence, compared to Typhimurium (Figure 28). These metabolites are commonly found in the gut, the ability to process these sugars can add a competitive edge against other intestinal microbiota. It has also been linked to increased production in response to intestinal inflammation in patients with microbiotal disorders such as Irritable Bowel Syndrome, supporting that it might help *Salmonella* compete against other microbes in the gut [207]. With that in mind the disruptions in this pathway in Choleraesuis SCA50 could explain why Choleraesuis is less effective at colonising the gut [137] compared to typhimurium. Whilst loss of this pathway is disadvantageous to growth, it may be advantageous to avoidance of the immune response. As we will see in the next section, there was no significance in the TraDIS data, supporting the theory that it is not an essential gene for pig colonisation. The loss of this function could be indicative that it is not required to colonise pigs (Figure 23).

A method of validation for the theory that pseudogene formation in Choleraesuis in sugar pathways leads to slower net replication in the gut because it is being out competed by other members of the gut flora could be to try growing Choleraesuis and Typhimurium on different mediums and measure their net replication. The relevant genes could be knocked out in typhimurium, or knocked-in in SCA50, to further test the hypothesis.



LEGEND

1.1.1.1	Present in LT2 and SCA50
1.1.1.1	Present in LT2, all orthologs are pseudogenes in SCA50
1.1.1.1	Present in LT2, no orthologs in SCA50
1.1.1.1	Present in LT2, functional and pseudogene orthologs in SCA50

Figure 28 KEGG pathway map of Pentose and Glucuronate. Green shows functional genes in Choleraesuis SCA50 and Typhimurium LT2. Red outline shows pseudogene formation in Choleraesuis SCA50 and Blue outline shows gene absence in Choleraesuis SCA50. Note that some green areas with a red outline, this occurs when this there is a version of a functional gene and a pseudogene.

3.2.3.5 Analysis of KEGG at the network level

The initial KEGG network in its entirety is enormous, it is complex and there is a skew of interactions between nodes. That is, some nodes have very many interactions and others have a single interaction. These kinds of network produce a hairball with thousands of very small clusters (Figure 29). This is a limitation of manmade networks, much like the internet there is a bias towards certain nodes. In order to be able to use the network for clustering analysis it needs to be pruned, this requires biological knowledge, pruning the wrong nodes would result in losing vital information. After pruning, some small groups of nodes became sub networks (Figure 30). We found that the clustering improved considerably with pruning, the clusters were bigger and fewer (4-19 nodes per cluster and 342 clusters) compared to the original network which had over 3000 clusters.

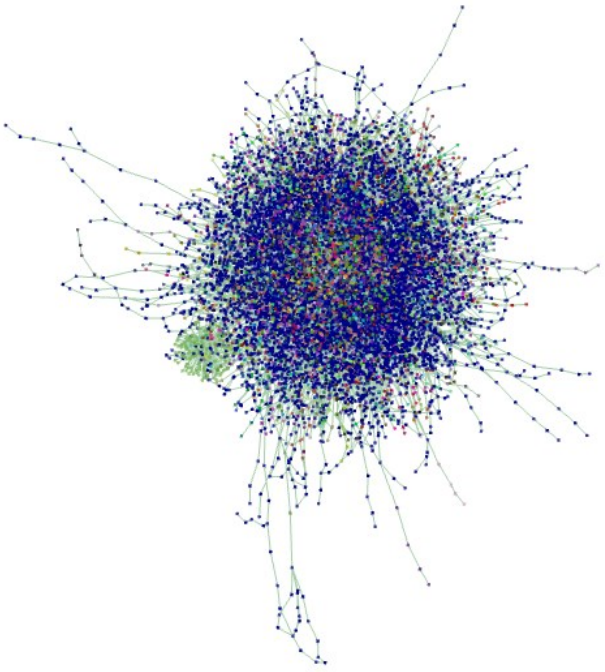


Figure 29 The entire KEGG network visualised in BioLayout, the different coloured nodes represent cluster group. This network clustered into 3410 clusters, most nodes did not group into a cluster (dark blue).

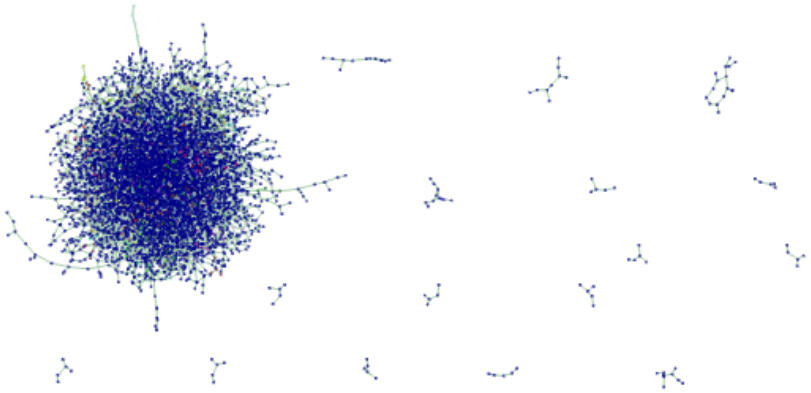


Figure 30 The KEGG network after pruning to remove ubiquitous compounds, sub-networks have formed ranging in size from 5-16 nodes. The clustering (groups coloured differently) consisted of 342 clusters with the number of nodes per cluster ranging from 4-19.

Further to this, extreme pruning gave a tree that formed reasonable clusters. The extreme pruning consisted of only keeping gene-gene interactions from a KEGG sub-

network (Microbial metabolism in diverse environments). This sub-network was used because it was in a more parsable format and reactions were not nodes (see section 3.2.3.5.1). This made the network more manageable for a proof of concept as mapping pseudogenes had to be performed manually. Figure 31 (A) shows that it is possible to get reasonable clustering in the KEGG network but it needs to be performed on one of KEGG's sub-networks. BioLayout served as a good means of visualising pseudogenes in the context of a larger network (Figure 31 B), the ability to easily cluster the network also meant that pseudogene cluster assignment could easily be extracted and used for enrichment analysis (Table 15).

I believe that there is potential to expand the analysis here to the entire KEGG network but it requires time and resources beyond the scope of this project.

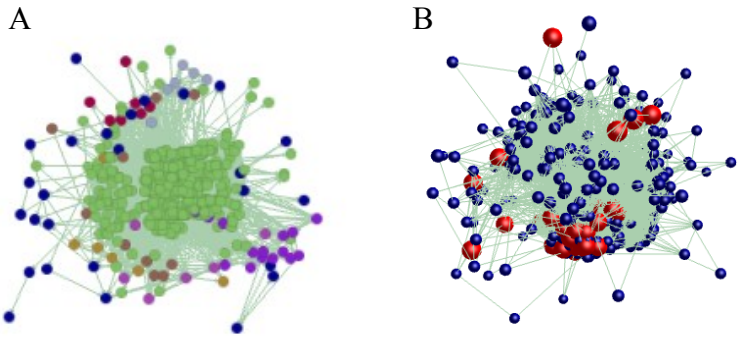


Figure 31 The extreme pruned sub-network of microbial metabolism in diverse environments. A shows MCL clustering. B shows the pseudogenes mapped to this mini network.

	total	expectation	observation	hypergeometric	fisher
Cluster01	415	28	21	0.999965	0.001025
Cluster04	7	0	3	0.036668	0.021375
Cluster02	15	1	4	0.036668	0.024445
Cluster03	10	1	3	0.042344	0.031758
Cluster07	6	0	1	0.436245	0.348996

Table 15 The enrichment analysis of the small sub-network for clusters with pseudogenes.

3.2.3.5.1 Difficulties with the analysis

One of the main difficulties with this analysis was computing power, the computer that this analysis was performed didn't not have the capacity to perform KEGG network scale analyses quickly; on occasion it took over twenty minutes just to open the file in BioLayout and another hour to perform the clustering. The KEGG network is large, it contains all of the compounds associated with every chemical reaction. This includes compounds like H^+ and ATP which are virtually ubiquitous to every chemical reaction. There are other networks in KEGG, these are smaller and are formatted to not include reactions as nodes. Further to this the genes are all assigned KEGG orthology IDs, making the node annotation more standardised. Table 16 and Table 17 give examples of the different format. In order to be able to make a version of the entire network that is in the more cluster friendly format all of the pathways' network files (KGML) could be downloaded individually and then combined. This would automatically link the nodes and interactions. The only limitation with this is that interactions the occur across two pathways might be omitted (Figure 32).

rn:R00305	-	AH2
rn:R00305	-	Gluconolactone
E1.14.11.9	-	rn:R02444_catalysis
rn:R01563	-	CO2
rn:R01563	-	NH3
rn:R01563	-	Sarcosine
rn:R03197	-	Coproporphyrinogen III
rn:R03197	-	CO2
2-Dehydro-D-gluconate	-	rn:R02658
rn:R01070	-	Glyceraldeh... 3-phosphate
rn:R01070	-	Glycerone phosphate
hemA	-	rn:R04109_catalysis
glycine for...ltransferase	-	rn:R02729_catalysis

Table 16 Part of the entire KEGG network showing that nodes are not annotated in a standard format, with some being truncated and a mixture of enzymes and gene names.

#koK01702_koK01703_koK01704	SEQUENTIAL_CATALYSIS	#koK00052
#koK01687	SEQUENTIAL_CATALYSIS	#koK00826
#cpdC00026	REACTS_WITH	#cpdC00407
#koK00263	SEQUENTIAL_CATALYSIS	#koK00053
#koK00052	METABOLIC_CATALYSIS	#cpdC00011
#EC14123	METABOLIC_CATALYSIS	#cpdC00003
#cpdC00011	REACTS_WITH	#cpdC00003
#koK01649	METABOLIC_CATALYSIS	#cpdC02504

Table 17 Part of a pathway network file, showing that it is formatted differently and that nodes can hold multiple genes.

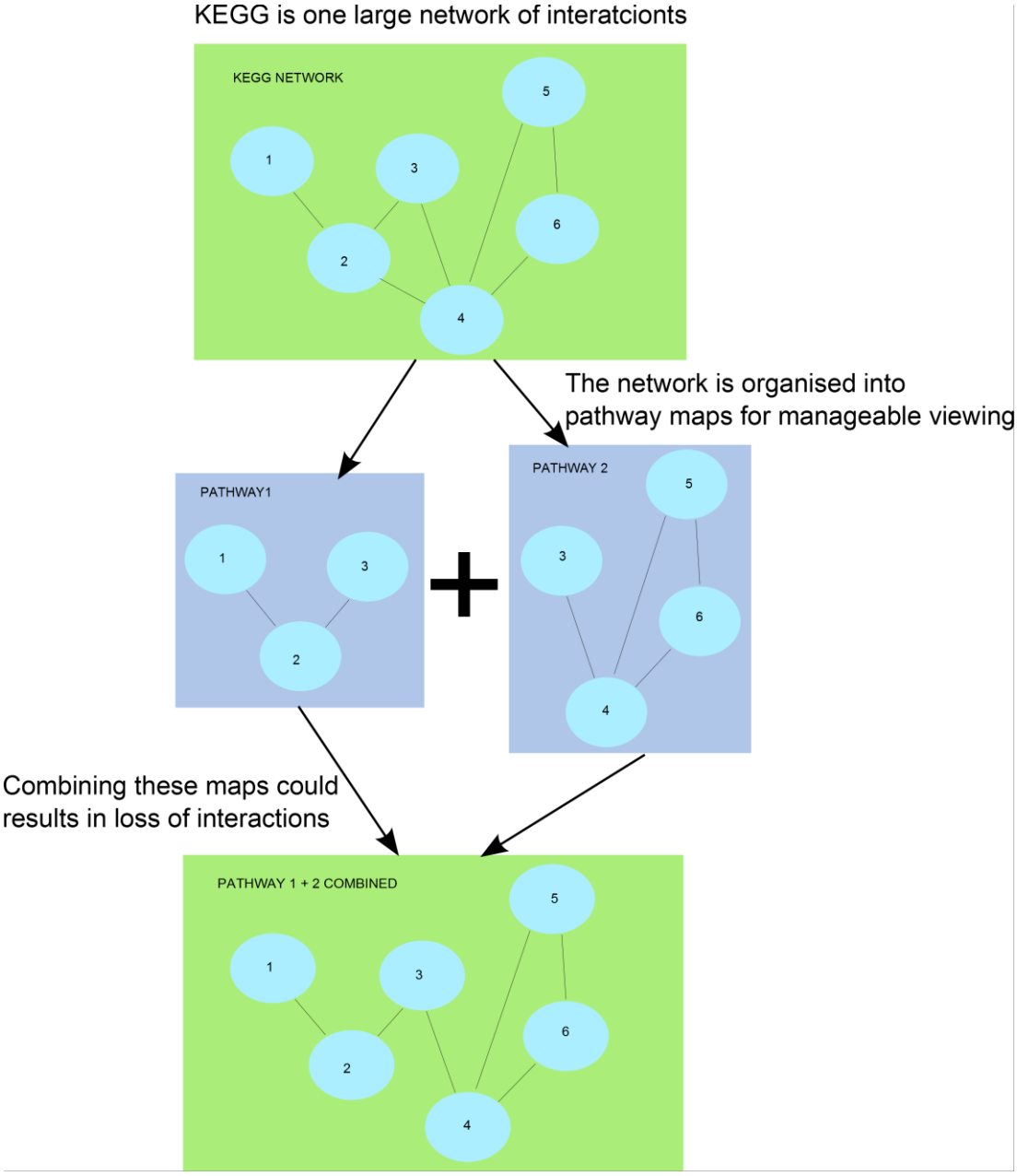


Figure 32 Shows the limitation of using combined data from pathways to infer the entire KEGG network, in this example the interaction between node 2 and 4 is lost.

3.3 TraDIS Analysis of Typhimurium SL1344

The TraDIS (Transposon directed insertion-site sequencing) data in this chapter is taken from Chaudhuri et al [208]. The authors studied *Salmonella enterica* serovar Typhimurium strain SL1344 in chickens, pigs and cattle. The importance of this study is that the authors are studying a pathogen in food-producing hosts that are of worldwide significance, rather than a model system.

The group produced a library consisting of over 7000 mutants, where each mutation was identified through DNA sequencing. They compared relative frequencies of mutants in pools before and after oral delivery into the host in question and the abundance of each mutation was measured as the cell count in culture. From this the authors were able to assign a relative fitness score, and associated p-value, to each mutant. The locations of the inserts were cross-referenced with the locations of genes, providing insight into the importance of each gene during the colonisation and infection of each of the three hosts. The methodology employed by Chaudhuri et al. is shown in Figure 33.

The group specifically looked for attenuating mutants, these are mutants that showed significantly decreased fitness relative to the wild type. They found enrichment in attenuation of pathogenicity islands and some species specific colonisation factors, for example, flagellar motility genes were only negatively attenuated in pigs.

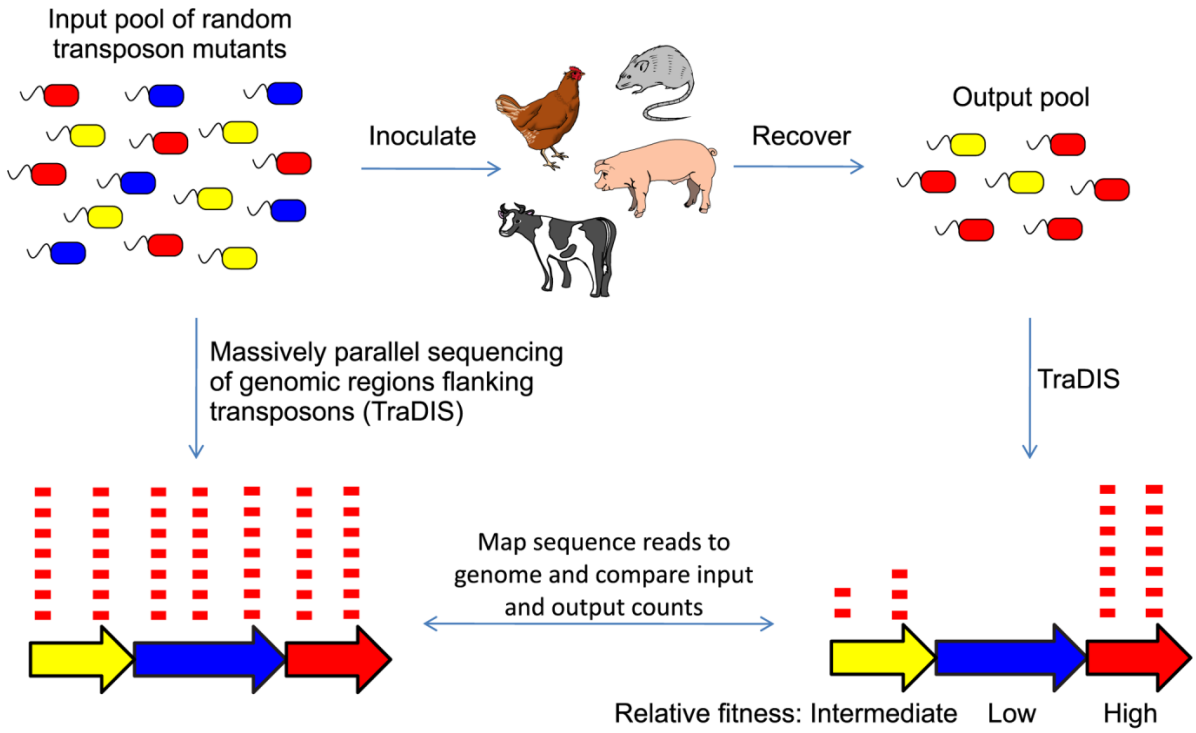


Figure 33 In vivo TraDIS methodology employed by Chaudhuri et al. (image taken from [208]).

The mutagenesis data produced in the Chaudhuri paper can indicate whether a gene is needed for infection. This section is divided into two parts. The first part counts the occurrence of each pathway associated with these seemingly essential (negatively selected) genes and performs enrichment analysis for selection. If any of the counts are drastically lower in one host than the others this may indicate that this pathway is not essential for infecting that particular host.

The second section maps our four serovars to the TraDIS data. By looking at patterns of pseudogene formation against the attenuation of mutations in a host generalist, it might be possible to find patterns between the mechanism of infection and the host. For example the data can be used to find non-essential genes for infecting a particular host systemically; if Gallinarum SG9 has a particular pseudogene whose ortholog in Typhimurium SL1344 is not negatively selected in chickens we can speculate that this pathway is required for the colonisation of chickens, systemically or enterically. Conversely, if a particular pseudogene is present in Gallinarum SG9 but is negatively selected in chickens in the TraDIS data this suggests that the gene is linked to enteric infection but not essential for systemic infection.

It is worth noting that at the time of this analysis Typhimurium SL1344 was not annotated, the TraDIS data was based on open reading frames. Repeating the analysis now may produce better mapping due to the available SL1344 annotation and the fact that some of the serovars are now available on KEGG, with KO assigned.

3.3.1 Attenuation score

The attenuation score shows whether more or less mutants colonise the host compared to the wild type's colonisation number. Therefore, a mutation with a negative score implies that the gene is essential for infection of that host. Conversely, if a mutant has a positive score we hypothesise that the gene is not essential for colonisation, and possibly losing the gene confers an advantage.

3.3.2 Methods

The data for this chapter was analysed by the authors, and provided to me as a spreadsheet. The analyses presented in this thesis represent my work, although all work up to and including the generation of the original table were carried out by the authors of the paper.

3.3.2.1 Pathway analysis

Initially the pathways allocated to each serovar in section 3.2.1.2 were used. However, when the annotation of Typhimurium SL1344 became publicly available along with the orthology file in KEGG this section was repeated with SL1344 as the reference rather than LT2. The fact that the TraDIS analysis was performed on SL1344 meant that rather than using homology methods to transfer the pathways a new script was made. ***`map_mutations.pl`*** gets the genes assigned to each mutation point and then map the pathways across from the KEGG orthology file (Appendix A: map_mutations.pl).

Enrichment analysis was then performed on this data set (as described in 3.2.1.4) with the population counts as all of the pathways that correspond to a mutated gene in the TraDIS data. Conversely, the sample group was made using ***`TraDIS_sig_genes.pl`***, this script goes through each mutation and counts the pathway if the p-value was below 0.05 and the fold change was below the accepted threshold for the host defined by Chaudhuri et al. [208] (Appendix A: TraDIS_sig_genes.pl).

3.3.2.2 Pseudogene ortholog analysis

Each mutant in the TraDIS data was mapped to the corresponding genes from the four sequenced serovars in Chapter 2. This was performed as in section 3.2.1.2, using ***`reciprocal_fasta.pl.`*** The TraDIS table was augmented with these orthologs.

Some genes did not map across even though there was an ortholog in Typhimurium SL1344. The genes which showed significant negative selection and did not appear to have any orthologs were mapped across manually.

If the corresponding gene was actually a pseudogene this was recorded. The 'FC' columns are shaded with a gradient of colour going from red for positive fold change to green for negative fold change, the threshold for the individual host was applied as a centre point for the gradient (-2,-3,-3 in chicks, calves and pigs respectively). The p-value column has cells coloured in green if the value is significant (below 0.05). For the full TraDIS results with orthologs and alternative colouring see Appendix A: Tradis_with_orthologs.xlsx.

3.3.3 Results

This section initially describes the TraDIS results that are common across all hosts, it then goes on to look at the individual results for each host (Appendix A: Tradis_summary.xlsx)

3.3.3.1 Results across all four hosts

There were 6 gene mutations with significant negative selection across all hosts where the corresponding serovars (bar Typhimurium 4/74) either had a pseudogene or the gene was completely absent (Table 18).

Two pathways were significantly enriched across all hosts after adjusting the p-values, 'Bacterial Secretion' and 'Lipopolysaccharide biosynthesis'. Three other enriched pathways were shared across all hosts' top ten pathways, namely 'Streptomycin biosynthesis', 'Polypeptide sugar unit biosynthesis and 'Lipoic acid metabolism' (Table 20, Table 22 and Table 24)

Table 18 Genes which show significant negative selection across all hosts and show pseudogene/absence (labelled as P/A in ortholog status respectively) in our host restricted/specific serovars

SL1344 Locus	STM 4/74 Locus	Product	KEGG Pathway	Serovar Ortholog Status		
				Chol	Dub	Gall
SL1344_3228	STM474_3412	Putative phosphotransferase system fructose-specific component IIB	00051 Fructose and mannose metabolism 01100 Metabolic pathways 02060 Phosphotransferase system (PTS)	A	A	A
SL1344_4149	STM474_4407	Putative phage tail sheath protein	None	P	A	A
SL1344_2166	STM474_2278	Galactoside transport atp-binding protein mgla	ABC transporters	P	P	P
SL1344_1930	STM474_2034	Tail protein	None	A	P	A
SL1344_3227	STM474_3411	Putative fructose-1-phosphate kinase	00051 Fructose and mannose metabolism	A	A	A
SL1344_0956	STM474_1007	Putative prophage protein	None	A	A	A

There were two pseudogenes in Typhimurium 4/74 that showed significant negative selection in the TraDIS data across all hosts. Firstly, SL1344_1437 and its ortholog, STM474_1517, are both annotated as pseudogenes. The other pseudogene was STM474_4286 whose ortholog, SL1344_4051, was not annotated as a pseudogene.

STM474_1517 and its ortholog SL1344_1437 are both pseudogenes and closer inspection shows they have identical sequences. Their description is as follows:

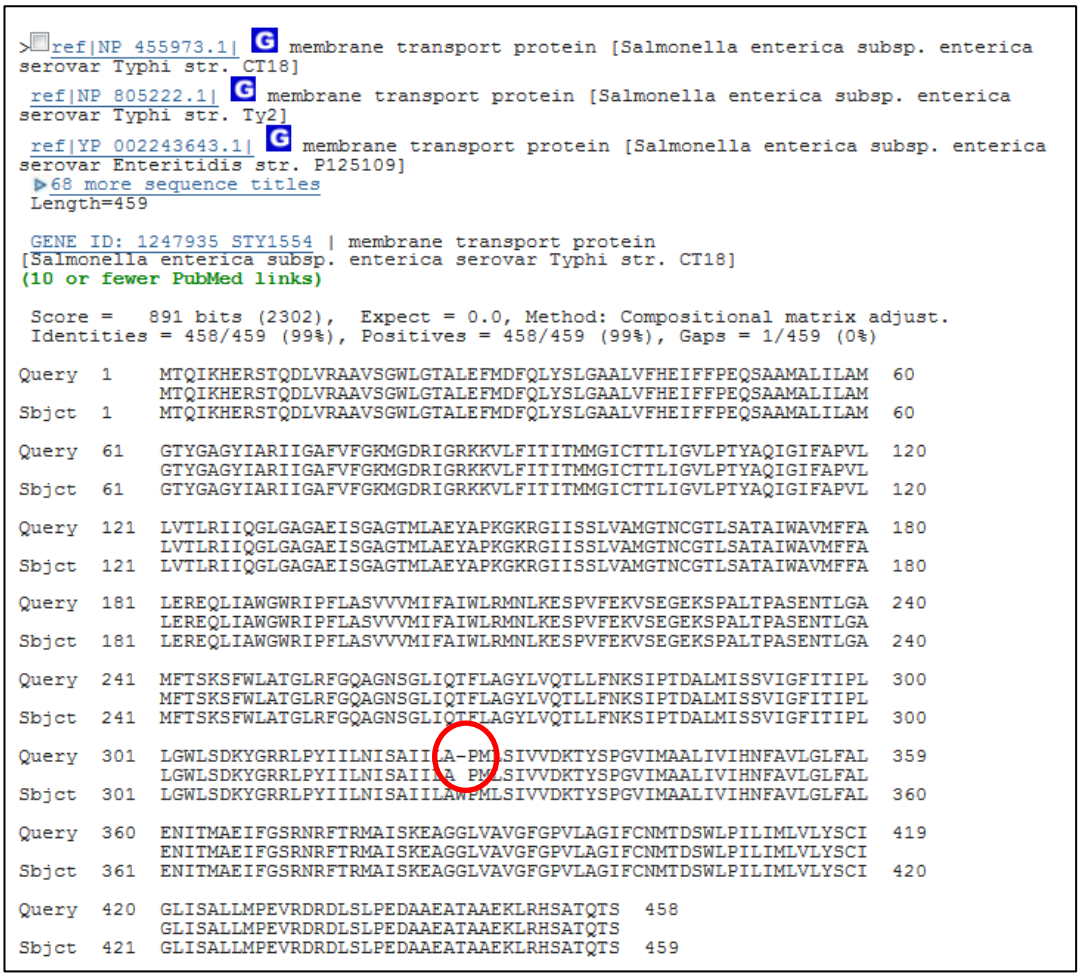
In-frame stop codon against amino acid: W; 'TGA' may be suppressed; corresponds to NP_455973.1 putative membrane transport protein; Pfam domain (MFS_1) truncated.

Performing BLAST against nr shows that the annotation is correct (Figure 34) there is a stop codon (TGA) where there should be a W (Tryptophan TGG) and this stop codon disrupts the domain.

The other pseudogene in 4/74 with negative selection was STM474_4286, its description is:

In-frame stop codon against 'W'; stop codon 'TAG';corresponds to NP_462983.1 putative inner membrane protein/MscS Mechanosensitive ion channel; Pfam domain(MS_channel) truncated.

The ortholog is SL1344_4051, this is not annotated as a pseudogene. Comparing the genomes in ACT showed that 4/74's gene is larger, spanning a stop codon (Figure 35). A BLAST search of the 4/74 sequence against nr shows that the hits are the same size but have a Tryptophan rather than a stop codon in the sequence (Figure 36). As the sequence across both SL1344 and 4/74 is identical the larger 4/74 annotation was accepted suggesting that SL1344_4051 is also a pseudogene with significant negative selection.



>ref|NP_455973.1| membrane transport protein [Salmonella enterica subsp. enterica serovar Typhi str. CT18]
ref|NP_805222.1| membrane transport protein [Salmonella enterica subsp. enterica serovar Typhi str. Ty2]
ref|YP_002243643.1| membrane transport protein [Salmonella enterica subsp. enterica serovar Enteritidis str. P125109]
▶68 more sequence titles
Length=459

GENE ID: 1247935 STY1554 | membrane transport protein
[Salmonella enterica subsp. enterica serovar Typhi str. CT18]
(10 or fewer PubMed links)

Score = 891 bits (2302), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 458/459 (99%), Positives = 458/459 (99%), Gaps = 1/459 (0%)

```
Query  1    MTQIKHERSTQDLVRAAVSGWLGTALEFMDFQLYSLGAALVFHEIFFPEQSAAMALILAM  60
            MTQIKHERSTQDLVRAAVSGWLGTALEFMDFQLYSLGAALVFHEIFFPEQSAAMALILAM
Sbjct  1    MTQIKHERSTQDLVRAAVSGWLGTALEFMDFQLYSLGAALVFHEIFFPEQSAAMALILAM  60

Query  61   GTYGAGYIARIIGAFVFGKMGDRIGRKKVLFITITMMGICTTLIGVLPTYAQIGIFAPVL  120
            GTYGAGYIARIIGAFVFGKMGDRIGRKKVLFITITMMGICTTLIGVLPTYAQIGIFAPVL
Sbjct  61   GTYGAGYIARIIGAFVFGKMGDRIGRKKVLFITITMMGICTTLIGVLPTYAQIGIFAPVL  120

Query  121  LVTLRIIQGLGAGAEISGAGTMLAEYAPKGKRGIISSLVAMGTNCGTLSATAIWAVMFFA  180
            LVTLRIIQGLGAGAEISGAGTMLAEYAPKGKRGIISSLVAMGTNCGTLSATAIWAVMFFA
Sbjct  121  LVTLRIIQGLGAGAEISGAGTMLAEYAPKGKRGIISSLVAMGTNCGTLSATAIWAVMFFA  180

Query  181  LEREQLIAWGWRIPFLASVVVMIFAIWLRMNLKESPVFEKVSEGEKSPALTPASENTLGA  240
            LEREQLIAWGWRIPFLASVVVMIFAIWLRMNLKESPVFEKVSEGEKSPALTPASENTLGA
Sbjct  181  LEREQLIAWGWRIPFLASVVVMIFAIWLRMNLKESPVFEKVSEGEKSPALTPASENTLGA  240

Query  241  MFTSKSFWLATGLRFGQAGNSGLIQTFLAGYLVQTLLFNKSIPTDALMISSVIGFITIPL  300
            MFTSKSFWLATGLRFGQAGNSGLIQTFLAGYLVQTLLFNKSIPTDALMISSVIGFITIPL
Sbjct  241  MFTSKSFWLATGLRFGQAGNSGLIQTFLAGYLVQTLLFNKSIPTDALMISSVIGFITIPL  300

Query  301  LGWLSDKYGRRLPYIILNISAIILA-PMLSIVVDKTYSPGVIMAALIVIHNFAVLGLFAL  359
            LGWLSDKYGRRLPYIILNISAIILA PMLSIVVDKTYSPGVIMAALIVIHNFAVLGLFAL
Sbjct  301  LGWLSDKYGRRLPYIILNISAIILAWPMLSIVVDKTYSPGVIMAALIVIHNFAVLGLFAL  360

Query  360  ENITMAEIFGSRNRFTRMAISKEAGGLVAVGFGPVLAGIFCNMTDSWLPILIMLVLYSCI  419
            ENITMAEIFGSRNRFTRMAISKEAGGLVAVGFGPVLAGIFCNMTDSWLPILIMLVLYSCI
Sbjct  361  ENITMAEIFGSRNRFTRMAISKEAGGLVAVGFGPVLAGIFCNMTDSWLPILIMLVLYSCI  420

Query  420  GLISALLMPEVRDRDLSLPEDAAEATAAEKLRHSATQTS  458
            GLISALLMPEVRDRDLSLPEDAAEATAAEKLRHSATQTS
Sbjct  421  GLISALLMPEVRDRDLSLPEDAAEATAAEKLRHSATQTS  459
```

Figure 34 Blast results of the pseudogene SL1344_1437 against nr, shows that the stop codon, represented by a '-' and circled in red is in the middle of domain and confirms this gene as a pseudogene.

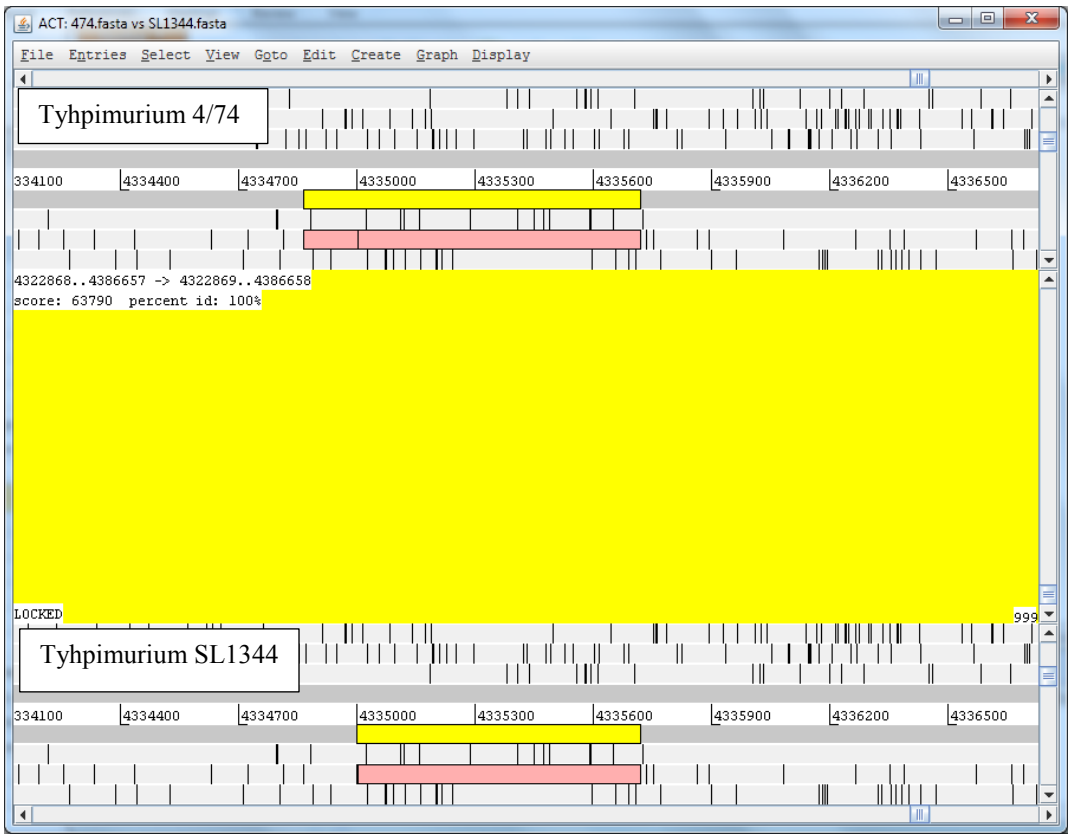


Figure 35 Comparison of 4/74 and SL1344 using BLAST shows the alignment for the region of pseudogene STM474_4286 (Typhimurium 4/74) and non-pseudogene SL1344_4051 (Typhimurium SL1344). The 4/74 pseudogene STM474_4286 is larger and spans a stop codon. SL1344_4051 stops at the codon which is spanned by STM474_4286.

```
>ref|NP_462983.1| inner membrane protein [Salmonella enterica subsp. enterica serovar
Typhimurium str. LT2]
 ref|ZP_02572779.1|  MscS Mechanosensitive ion channel [Salmonella enterica subsp.
enterica serovar 4,[5],12:i:- str. CVM23701]
 ref|ZP_03165868.1|  MscS Mechanosensitive ion channel [Salmonella enterica subsp.
enterica serovar Saintpaul str. SARA23]
 ▶13 more sequence titles
 Length=285

 GENE ID: 1255629 STM4102 | inner membrane protein [Salmonella typhimurium LT2]
(10 or fewer PubMed links)

 Score =   564 bits (1453),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 284/285 (99%), Positives = 284/285 (99%), Gaps = 1/285 (0%)

Query  1    MEGFELFPKIKGAIKWMAEHSDSVIHFGWNVVAAIILLFIGKLIARLLSRGLEKLLLRRQ  60
            MEGFELFPKIKGAIKWMAEHSDSVIHFGWNVVAAIILLFIGKLIARLLSRGLEKLLLRRQ
Sbjct  1    MEGFELFPKIKGAIKWMAEHSDSVIHFGWNVVAAIILLFIGKLIARLLSRGLEKLLLRRQ  60

Query  61   VDATIVHFFSALVRYITIAFTAVAALGRVGIETSSIIAVIGAAGLAIGLALQGSLSNFAA  120
            VDATIVHFFSALVRYITIAFTAVAALGRVGIETSSIIAVIGAAGLAIGLALQGSLSNFAA
Sbjct  61   VDATIVHFFSALVRYITIAFTAVAALGRVGIETSSIIAVIGAAGLAIGLALQGSLSNFAA  120

Query  121  GVLLVSLRPFRAGEIVQIGLVIGTVEKVHIFSTTLLTADSKEVVIPNGKIIADNIINYSR  180
            GVLLVSLRPFRAGEIVQIGLVIGTVEKVHIFSTTLLTADSKEVVIPNGKIIADNIINYSR
Sbjct  121  GVLLVSLRPFRAGEIVQIGLVIGTVEKVHIFSTTLLTADSKEVVIPNGKIIADNIINYSR  180

Query  181  HPYRRIDLIIGVDYQSRIADVKNVIHRIIEQDHRIDKTRDITVRLGELAPSSLNFYVRV-  239
            HPYRRIDLIIGVDYQSRIADVKNVIHRIIEQDHRIDKTRDITVRLGELAPSSLNFYVRV
Sbjct  181  HPYRRIDLIIGVDYQSRIADVKNVIHRIIEQDHRIDKTRDITVRLGELAPSSLNFYVRVW  240

Query  240  VPNAQYWSTYYDLLENIKEAMDENGINIPYPRMDVRVENVKSITP  284
            VPNAQYWSTYYDLLENIKEAMDENGINIPYPRMDVRVENVKSITP
Sbjct  241  VPNAQYWSTYYDLLENIKEAMDENGINIPYPRMDVRVENVKSITP  285
```

Figure 36 Blast results of the pseudogene STM474_4286 against nr, shows that the stop codon, represented by a '-' and circled in red is in the middle of domain and confirms this gene as a pseudogene. It also shows that the sequences it hits span across the stop codon.

3.3.3.2 Calf

Pseudogene analysis of the TraDIS data returned 4 calf specific genes. That is genes that are negatively selected only in calves and are only functional in Dublin SD3246 (and Typhimurium 4/74) (Table 19). Two of them are labelled as phage protein. One gene is linked to the 'Methane metabolism pathway' and the other is a large repeat protein.

Fisher's exact test results for negative selection in Calves showed two significant results after adjusting the p-values, 'Bacterial secretion system' and 'Lipopolysaccharide biosynthesis' (Table 20). 'Citrate cycle' was significant before adjusting the p-value with 11 negatively selected genes out of 20 (the expected frequency was 5).

Positive selection tests showed negative enrichment in 'Bacterial secretion system' which means that there were less positively selected mutations in 'Bacterial secretion' than expected (Eight expected, two observed), this supports the negative selection results in calves. Other negatively enriched mutations for positive selection included 'Nucleotide excision repair'. Conversely, the pathway 'Folate biosynthesis' and 'Nicotinate and nicotinamide metabolism' showed positive enrichment for positive selection in calves (Table 21).

Table 19 The four calf specific negatively selected genes (from TraDIS data for SL1344) which are only functional in Dublin 3246 and show pseudogene/absence in the other host specific/restricted serovars. F – functional, A – absent, S – Selenocysteine readthrough.

SL1344 Locus	SD3246 Locus	Product	KEGG Pathway	Serovar Ortholog Status		
				Typh	Chol	Gall
SL1344_1486	SD3246_1735	Na+/H+ antiporter NhaC	Methane metabolism	F	A	P
SL1344_2704	SD3246_2962	Hypothetical bacteriophage protein	None	F	A	A
SL1344_2709	SD3246_2967	Bacteriophage integrase	ABC transporters	F	A	A
SL1344_2661	SD3246_2921	VCBS repeat-containing protein and Large repetitive protein bapA	None	F	S	P

Table 20 Top 10 negatively selected pathways in calves. Pathways are organised into those positive enrichment (more observed than expect) and negative enrichment (less observed than expected), green and pink cells respectively. Red text indicates pathways that are significantly negatively selected

Pathway	Description	Population	Expected	Observed	Hypergeometric	Fisher's	Adjusted Fisher's
P03070	Bacterial secretion system	20	5	14	0.000864	8.64E-06	0.000864
p00540	Lipopolysaccharide biosynthesis	7	2	6	0.040965	0.000819	0.040965
p00020	Citrate cycle (TCA cycle)	20	5	11	0.060131	0.001804	0.060131
p00521	Streptomycin biosynthesis	9	2	6	0.143674	0.006411	0.160279
p00190	Oxidative phosphorylation	29	7	13	0.143674	0.011365	0.175344
p00523	Polyketide sugar unit biosynthesis	5	1	4	0.175344	0.011481	0.175344
p00785	Lipoic acid metabolism	3	1	3	0.175344	0.012274	0.175344
p05132	Salmonella Infection	11	3	6	0.288654	0.023092	0.288654
p05100	Bacterial Invasion of Epithelial Cells	6	1	4	0.313537	0.028218	0.313537
p03010	Ribosome	4	1	3	0.406721	0.040672	0.406721

Table 21 Top 10 positively selected pathways in calves Pathways are organised into those positive enrichment (more observed than expect) and negative enrichment (less observed than expected), green and pink cells respectively. Red text indicates pathways that are significantly positively selected

Pathway	Description	Population	Expected	Observed	Hypergeometric	Fisher's	Adjusted Fisher's
p03070	Bacterial secretion system	20	8	2	0.999594	0.004574	0.457362
p00790	Folate biosynthesis	6	2	5	0.042451	0.042451	1
p03420	Nucleotide excision repair	7	3	0	1	0.045702	1
p00500	Starch and sucrose metabolism	25	10	15	0.03578	0.060594	1
p00400	Phenylalanine, tyrosine and tryptophan biosynthesis	20	8	4	0.986431	0.066853	1
p03030	DNA replication	5	2	0	1	0.084941	1
p00640	Propanoate metabolism	14	6	9	0.060846	0.096731	1
p00010	Glycolysis / Gluconeogenesis	25	10	6	0.975186	0.100385	1
p00760	Nicotinate and nicotinamide metabolism	11	4	7	0.10333	0.13088	1
p00240	Pyrimidine metabolism	25	10	14	0.081424	0.146029	1

3.3.3.3 Chick

There were no genes that showed chick specific negative selection whilst having functional orthologs in Gallinarum SG9 and non-functional orthologs/absence in Dublin SD3246 and Choleraesuis SCA50.

The enrichment analysis revealed three pathways that were significant after adjusting the p-value. As with the Calf enrichment analysis 'Lipopolysaccharide biosynthesis' and 'Bacterial secretion system' were significant. In addition other pathways including Streptomycin biosynthesis were significant after adjustment. Two negatively enriched pathways for negative selection were in the top ten enriched pathways for chicks. Negative enrichment in this case means that number of negatively selected mutations for a given pathway were less than expected. The two pathways were 'Flagellar Assembly' (4 expected, zero observed) and 'Two-Component system' (24 expected, 16 observed).

The positive selection enrichment analysis concurs with the above, showing enrichment for 'Flagella assembly' (Table 23). 'Arginine and proline' metabolism is also significantly enriched. This concurs with section 3.2.2.2.3 which shows Gallinarum specific pseudogene formation and higher observed pseudogene formation than predicted in both these pathways.

Table 22 Top 10 negatively selected pathways in chicks. Pathways are organised into those positive enrichment (more observed than expect) and negative enrichment (less observed than expected), green and pink cells respectively. Red text indicates pathways that are significantly negatively selected

Pathway	Description	Population	Expected	Observed	Hypergeometric	Fisher's	Adjusted Fisher's
p00540	Lipopolysaccharide biosynthesis	7	2	7	0.004247	4.25E-05	0.004247304
p03070	Bacterial secretion system	20	5	13	0.004773	9.55E-05	0.004773002
p00521	Streptomycin biosynthesis	9	2	7	0.032285	0.000969	0.032284944
p02040	Flagellar assembly	18	4	0	1	0.010299	0.171210444
p03018	RNA degradation	12	3	7	0.17121	0.010811	0.171210444
p00523	Polyketide sugar unit biosynthesis	5	1	4	0.17121	0.013182	0.171210444
p00785	Lipoic acid metabolism	3	1	3	0.17121	0.013697	0.171210444
p03060	Protein export	3	1	3	0.17121	0.013697	0.171210444
p01110	Biosynthesis of secondary metabolites	190	46	58	0.17121	0.019843	0.220482497
p02020	Two-component system	102	24	16	1	0.035383	0.346405917

Table 23 Top 10 positively selected pathways in chicks. Pathways are organised into those positive enrichment (more observed than expect) and negative enrichment (less observed than expected), green and pink cells respectively. Red text indicates pathways that are significantly positively selected

Pathway	Description	Population	Expected	Observed	Hypergeometric	Fisher's	Adjusted Fisher's
p02040	Flagellar assembly	18	3	8	0.004606	0.004606	0.271752
p00330	Arginine and proline metabolism	26	4	10	0.005435	0.005435	0.271752
p00400	Phenylalanine, tyrosine and tryptophan biosynthesis	20	3	0	1	0.059588	1
p01120	Microbial metabolism in diverse environments	121	20	13	0.980817	0.064286	1
p00600	Sphingolipid metabolism	3	0	2	0.072715	0.072715	1
p00340	Histidine metabolism	11	2	4	0.09209	0.09209	1
p05132	Salmonella Infection	11	2	4	0.09209	0.09209	1
p02010	ABC transporters	89	15	9	0.975474	0.096154	1
p00240	Pyrimidine metabolism	25	4	1	0.989906	0.102462	1
p00562	Inositol phosphate metabolism	8	1	3	0.131589	0.131589	1

3.3.3.4 Pig

There were no mutations in the TraDIS data that showed pig specific negative selection and loss of function in Gallinarum SG9 and Dublin SD3246 orthologous genes.

The only two enriched negatively selected pathways in the TraDIS data for pigs (after adjusting p-value) were 'Bacterial secretion system' and 'Lipopolysaccharide metabolism', which are also significantly negatively selected in calves and chicks (Table 24). Of the top ten enriched pathways for negative selection, the following were unique to pigs, 'Folate biosynthesis', 'Chloroalkane and chloroalkene degradation' and 'Naphthalene degradation' (Table 24).

There were three enriched pathways for positive selection, 'Lipopolysaccharide biosynthesis', 'Nitrogen metabolism' and 'Alanine, aspartate and glutamate metabolism' (Table 25
). The positive selection of 'Lipopolysaccharide biosynthesis' is contrary to the significant negative selection in table (Table 24)

Table 24 Top 11 negatively selected pathways in pigs. Pathways are organised into those positive enrichment (more observed than expect) and negative enrichment (less observed than expected), green and pink cells respectively. Red text indicates pathways that are significantly negatively selected

Pathway	Description	Population	Expected	Observed	Hypergeometric	Fisher's	Adjusted Fisher's
p03070	**Bacterial secretion system**	20	4	15	3.42E-05	3.42E-07	3.42E-05
p00540	**Lipopolysaccharide biosynthesis**	7	2	6	0.027007	0.00054	0.027007
p00521	Streptomycin biosynthesis	9	2	6	0.141426	0.00437	0.141426
p03018	RNA degradation	12	3	7	0.141426	0.005734	0.141426
p00523	Polyketide sugar unit biosynthesis	5	1	4	0.141426	0.008748	0.141426
p00785	Lipoic acid metabolism	3	1	3	0.141426	0.0099	0.141426
p03060	Protein export	3	1	3	0.141426	0.0099	0.141426
p00790	Folate biosynthesis	6	1	4	0.272881	0.021831	0.272881
p00190	Oxidative phosphorylation	29	6	11	0.341871	0.03784	0.420442
p00625	Chloroalkane and chloroalkene degradation	2	0	2	0.421027	0.046313	0.421027
p00626	Naphthalene degradation	2	0	2	0.421027	0.046313	0.421027

Table 25 Top 10 positively selected pathways in pigs. Pathways are organised into those positive enrichment (more observed than expect) and negative enrichment (less observed than expected), green and pink cells respectively. Red text indicates pathways that are significantly positively selected

Pathway	Description	Population	Expected	Observed	Hypergeometric	Fisher's	Adjusted Fisher's
p00540	Lipopolysaccharide biosynthesis	7	0	2	0.022938	0.022938	1
p00910	Nitrogen metabolism	35	1	4	0.031348	0.031348	1
p00250	Alanine, aspartate and glutamate metabolism	21	1	3	0.034892	0.034892	1
p00190	Oxidative phosphorylation	29	1	3	0.079215	0.079215	1
p00640	Propanoate metabolism	14	0	2	0.085213	0.085213	1
p00030	Pentose phosphate pathway	22	1	2	0.181939	0.181939	1
p00240	Pyrimidine metabolism	25	1	2	0.221629	0.221629	1
p00630	Glyoxylate and dicarboxylate metabolism	25	1	2	0.221629	0.221629	1
p00330	Arginine and proline metabolism	26	1	2	0.235054	0.235054	1
p02020	Two-component system	102	4	6	0.142241	0.245001	1

3.3.4 Discussion

The negative selection across all three hosts for 'Bacterial secretion' and 'Lipopolysaccharide biosynthesis' is consistent with *Salmonella* biology. The lipopolysaccharide layer and bacterial secretion system are both essential components of gram negative bacteria allowing them to interact with the outside world, in terms of infection, reproduction and adherence for example.

It is worth noting that one limitation of the TraDIS data is that there are many different mutations in a population, perhaps one mutation that would normally be viable in its own right is being outcompeted by something that is far more viable. One way to overcome this is to perform targeted mutagenesis on genes which are of specific interest after performing the TraDIS analysis.

3.3.4.1 Fructose in the role of gut colonisation

Pseudogene formation in 'Fructose and mannose metabolism' is seen across all serovars except Typhimurium 4/74 (Table 13). Coupled with this there is negative selection across all hosts in two genes associated with this pathway (Table 18). These genes are actually absent in the other serovars, meaning that they are not orthologous to the pseudogenes recorded in Table 13 Typhimurium had very few pseudogenes so the comparison seemed a little unfair, although this fact itself is indicative of its adaptation to multiple hosts.

Fructo-oligosaccharides can be found in the gut of many different animals as there are high levels of fructose based compounds in fruit, vegetables and grains. Fructo-oligosaccharides have been linked to competitive exclusion in gut microbiota [209] and the metabolism for these compounds is commonly found in bacteria associated with the gastrointestinal tract [210].

There have also been associations made with pathogenicity where fructose presence causes a decrease in salmonella resistance in rats [211] and can act as a trigger for the virulence gene program in enterohaemorrhagic Escherichia coli (EHEC) [212].

These findings give rise to some questions. Firstly, is the apparent attrition of this pathway in our host adapted serovars due to fructose and mannose being catabolised by other routes? We know that the genes in this pathway are significantly attenuated in Typhimurium, what would happen if these genes were reinstated in the other serovars? This would be especially interesting in Choleraesuis SCA50 because this is known to replicate more slowly in the gut compared to Typhimurium [137], could this metabolism disruption be causative?

3.3.4.2 The loss of motility is linked to increased pathogenicity in chickens

Gallinarum shows high gene attrition in pathways linked to motility such as 'cell motility' and 'flagellar assembly', compared to the other serovars and according to Enrichment analysis Figure 22 and Figure 25 respectively.

It is known that Gallinarum as a serovar is non-motile and this has been linked to pseudogene formation in flagella genes [213]. The Enrichment analysis of the TraDIS data shows that a host generalist like Typhimurium strain SL1344 might actually gain an advantage by losing genes in these pathways, (Table 22 and Table 23).

The pseudogene and TraDIS data presented here clearly correlates well with what we already know about *Salmonella* serovar gallinarum, in that it is non-motile. The data highly suggest that being non-motile confers an advantage during infection of the chicken. This was observed by Iqbal *et al.* who in the 2005 study found that mutating the fliM gene in Typhimurium 'showed an enhanced ability to establish systemic infection' in chickens [136]. The exact mode of action of this process is unknown, but it is encouraging that both genome analysis and TraDIS analysis support the hypothesis that loss of motility is advantageous. One possible explanation is that, by losing the ability to construct flagella, gallinarum is able to better avoid the host immune response, the flagellum being an obvious target. However, more work needs to be done to explain this, particularly why loss of the flagellum confers an advantage in chicks but not in calves or pigs.

The fact that these results concur with the literature demonstrates that this method of analysis provides scientifically meaningful results. To further examine the role of flagella in pathogenicity further mutation studies could be applied, looking at the effects of other serovars pathogenicity in chickens. For example, would knocking out these pathways in *Salmonella* serovar Dublin enable it to infect chickens?

3.4 Concluding Remarks

This chapter has taken post genomic data, namely pathways and mutagenesis and used them to explore enrichment in pseudogene formation and negative selection respectively. We see that both genome analysis and TraDIS analysis supports our knowledge of the biology of Salmonellae serovars, and we can begin to build up a picture of the pathways needed for different serovars to infect different hosts.

Making broad conclusions about phenotypes such as host specificity and pathogenicity are very difficult because they are complex traits. Further to this trying to do complete comparison between genomes is often hindered by different quality annotation. The rule of diminishing returns comes to mind. The current state of comparative genomics does mean that certain features will be missed. Ewan Birney explains in his blog some of the difficulties with large data sets stating that there will be errors, things will be missed and trying to compare work that has been done in two different labs won't be perfect [214]. For example annotations of Typhimurium SL1344 and STM4/74 should be nearly identical but they have some inconsistencies.

Trying to make sense of many large pieces of data is often cumbersome and can be difficult to interpret, methods of integrating many types of data into one hub would make such analyses easier.

Chapter Four

Design and implementation of GeneBook

Based on chapters 2 and 3 we have established that publically available data can contain errors and interpreting post genomic data is complex. Chapter 4 introduces GeneBook, a bacterial genome web explorer that integrates remote data sources into a single workspace.

The chapter has two main sections. The first, 4.2, explains the design and development of GeneBook; this is followed by a section demonstrating scenarios of how GeneBook can be used to get information above and beyond the bacterial genome resources that are currently available (section 4.3).

4.1 Aims

The main objective of this chapter is to produce a tool that allows users to synchronously analyse disparate datasets with a dynamic, intuitive interface. Taking their data from conception as a raw sequence and cultivating it into a fully annotated genome sitting at the hub of a web of unified heterogeneous data. It demonstrates how GeneBook incorporates data from different locations to see bacterial features in a more accurate and biologically meaningful way.

The aims of this chapter are:

- To design and implement an automated tool for a quick and rudimentary annotation of bacterial genomic sequences.
- Provide an interface for uploading genomic sequences
- To produce a database driven website that integrates heterogeneous data types into one central web-based focal point.
- Provide tools to link quantitative data to the genome annotation; ensuing a method of integrating and visualising data such as targeted mutagenesis and microarray.
- Provide tools for linking next generation sequencing (NGS) data with array data; integrating information about sequencing quality and expression studies, into the context of the genome.
- Give users, with potentially no programming skills, the means to link remote data sources and their own private data.
- To demonstrate how GeneBook:
    - Augments uninformative annotations with meaningful data.
    - Highlights omitted/missed connections in other databases.
    - Can be used to assess annotation inconsistencies.
    - Can assess pseudogene annotations.
    - Integrates real experimental data to build hypotheses around Salmonella biology.

4.2 Design and development of GeneBook

Section 4.2.1.1 uses the critical review in section 1.6 as a justification for a new method of storing and displaying bacterial genomic data, namely GeneBook. 4.2.1.2 describes the development of a lightweight, novel genome database which integrates disparate datasets. The methodology of using existing and making new web services is demonstrated in 4.2.1.3, showing the potential for accessing data from remote data sources. Finally, new methods of visualising generic quantitative data and NGS data are developed and discussed, with consideration of future developments as sequencing technologies advance.

4.2.1 Developing a system for managing disparate information resources

Given the diverse range of data available for bacteria, both genomic and post-genomic, and with the location of this data being spread across 10s if not 100s of different databases, we hypothesized that a new system that was lightweight and which used web-services to integrate data from disparate resources would enable scientists to gain a clearer picture of gene function. This hypothesis led to the development of GeneBook, a system for integrating diverse data types in one location using web-services. A diagram of how GeneBook works is shown in Figure 38. There is an underlying database storing only sequence and the location of genomic features, taken directly from the genome annotation. Data is served to users from an Apache web-server. To enable efficient integration, GeneBook uses the open-source web content-management system (CMS), Drupal. Both public and private data were integrated with the genome annotation using web-services, with widgets (or web-apps) developed inside the drupal CMS.

This section describes the database design and structure (4.2.1.1), the ethos behind the Drupal web CMS (4.2.1.2) and the web-services that were integrated (4.2.1.3)

4.2.1.1 Managing gene data

The resources described earlier are cumbersome in terms of their update status, mainly due to their static nature. Another consideration is if new data types are made, adding these to the schemas could be complex. Finally, people who access these resources are limited to publicly available data, and there is no option for them to

view their data alongside the data within the resource. A system with dynamic content, taken from remote data sources, would have no need for monthly updates as it would be as up-to-date as the primary source that the data is retrieved from. This section describes the design and development of a generic gene-centric database.

4.2.1.1.1 Database Design

GeneBook, described in this chapter's aims is based on a simple, lightweight, gene-centric database. The database only needs to carry the basic genomic information, from a genome format such as Embl or GenBank. Any extra information will be integrated dynamically using remote data sources. This will make the system less cumbersome than some of the existing bacterial genome resources.

There are a few relational database management systems (RDBMS) available at the Roslin Institute; Microsoft Access, Oracle, Firebird and MySQL. Both Microsoft Access and Oracle are not commonly used in the biological community as they are not open source. This means that there may be proprietary issues and little integration with other biological tools. Firebird and MySQL are both open source. MySQL was the chosen RDBMS because it is easily integrated with Apache, the web-server we will use, and PHP. Further to this, it is more universal, with a large community of users and is often the RDBMS of choice for biological database developers.

In terms of actually designing the database there are some schemas already available specifically for biological databases. In order to ascertain whether one of the available schemas would be suitable or if a schema should be made specifically for this project some of the major schemas were reviewed. The schema has to support MySQL and must be lightweight and flexible. Schemas such as the Genomics Unified Schema (GUS) [215] and Chado [216] are very comprehensive and flexible. This means that they can cater for a plethora of biological relationships. On the other hand, the extensive nature of these schemas means that they can take a long time to learn and simple queries can be overly complex to implement. Considering that the database needs to be lightweight, using one of these schemas would be excessive, both temporally and computationally. Other schemas like BioSQL [217] are more

lightweight and can be easily integrated with most biological programming languages (BioPerl, BioJava etc.). After examining the different schemas available it became apparent that the entities which they incorporate are far beyond the scope of this proposed database, which will focus on genomic features exclusively. It was decided that using any current schema would not be necessary, as the database will only need a few entities with easily identifiable relationships.

Figure 37 shows the final Entity Relationship Diagram (ERD. In terms of normalisation the gene entity should have been merged with the feature table. The justification for not doing this is that at the time of design it wasn't certain whether the other features would actually be used in the website as it was initially designed to be gene centric.

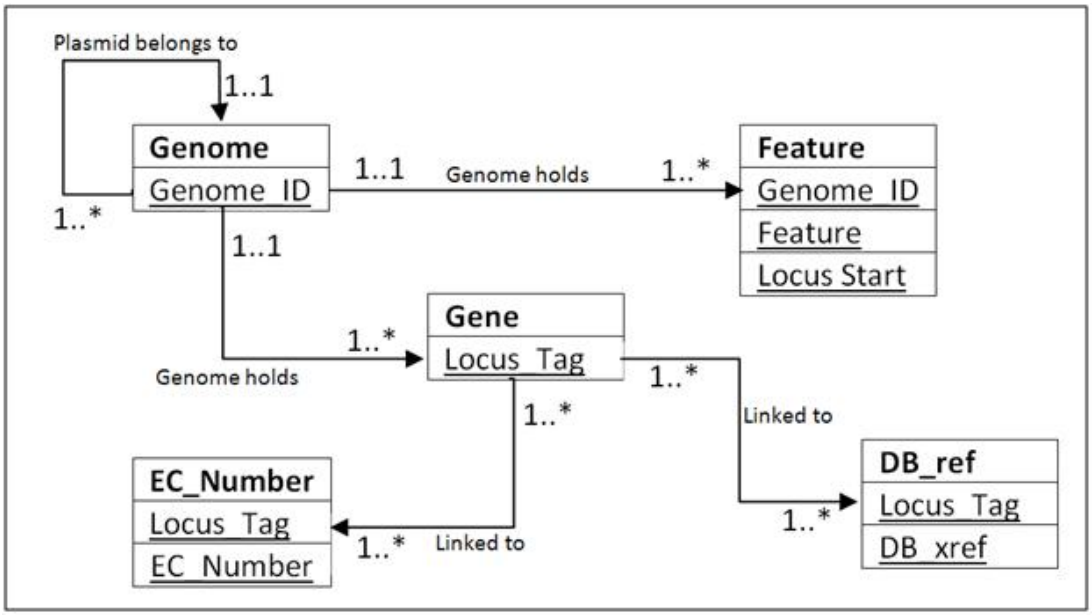


Figure 37 ERD for the GeneBook database, the fields showing the primary keys.

4.2.1.1.1.1 Future improvements

The database has been in use for over two years with virtually no problems. However, if I was to repeat the entity design aspect of this project I would put all the features together into one entity and store even less information than is currently stored. That is the entity should be limited to Genome ID, Feature (the type of feature), Locus tag (if available) and the location. As the genome sequences are stored locally the start and stop sites could be used to obtain the sequence which can be easily translated if necessary.

Within a year of its conception the database had to be migrated to a different server on a different site. The tool mysqldump dump was used for the transition. A limitation of the MySQL dump process caused the entities to no longer cascade by their foreign keys. This means that when information in an entity is deleted the data which used the deleted information is not deleted. So in terms of this database if an entry from the genome entity was deleted all of the features and genes would remain in the database referring to a genome which no longer exists This defies the database rules of normalisation and could result in orphan piece of data. It is possible to modify the database to include the cascading, and if time permits this task will be performed. Although the database no longer cascades this would only be problematic if multiple users could delete entries, currently only the administrator can delete entries.

4.2.1.1.2 Populating the database

All of the *Salmonella* genomes available at the NCBI at the time of populating (2009) were downloaded and processed for inclusion into the database

The process of parsing the genomes should have been straightforward. However, as there is no standard for bacterial genome annotation the quality and procedure was variable between genomes. This meant that parsing the GenBank files into a tab delimited file and dumping this into the database didn't work. Individual MySQL insert statements had to be constructed and dumped into the database in a batch.

RefSeq annotation was also used, which tries to standardise genome annotation, but even so there is variability in the way the annotations are described
EMBOSS's extractfeat was used to get the Coding Sequence (CDS) features from the genomes and these were then inserted into the database using SQL commands.

4.2.1.1.3 Novel Genomes

The availability of public data for each genome is skewed [2]. Newly sequenced genomes may not have much data in the public domain. This means that integrating some of the webservices which rely on locus tag will prove fruitless. As a future step part of the genome upload process could include an orthology step. That is the user can either include orthologs in their upload or orthologs could be calculate upon upload.

The ortholog locus tag could then be used to retrieve information should there be no information in the public databases for the actual locus tag. This might be limited by services which take a long time to return a failed result. Alternatively, the user could choose they want to see the original locus tag or its ortholog.

4.2.1.2 Developing GeneBook

During the development of GeneBook it became clear that a CMS was needed. This prevented time being wasted on developing user administration, web page aesthetics and scalability. The web interface section, 4.2.1.2.2, specifically describes the development of the web interface to the database described in 4.2.1.1.1. The functionality of the website and how this uses web 2.0 technologies to give a dynamic data driven website is demonstrated by showing the integration of remote and private data using webservices.

Webservices are an integral part of GeneBook, providing most of the information for each feature. They are defined in this section, using biological examples of in-house and remote data source integration.

The final product, GeneBook has wrappers which allow it to interact with remote websites, pulling and displaying the information into one space meaning that the user can see the feature and its data holistically.

4.2.1.2.1 Transition to a CMS

The first step was to design a simple web page which connects to the database. This was done using html to hold the content of the page, Cascading Style Sheets (CSS) to define the layout and style, and PHP to interface between the database and the web site.

Initially the interface was designed as a proof of concept, demonstrating that remote web services can be integrated into a web page on-the-fly. This interface was able to use web 2.0 to autopredict gene names and was able to show remote data on an *ad-hoc* basis. It had some drag and drop functionality but this was not maintained between sessions.

Implementation of a CMS handles user login and other data, it also records user activities and preferences. Most importantly it allows the project to focus on the scientific side of development rather than administration (CMS development, CSS modification for example). A new interface was developed to implement these features using Drupal (Figure 38). A CMS also gives scope for private user login, private workspaces and follows a logical directory framework providing a basis for webservice integration.

The Drupal [218] CMS was adopted because it is written in PHP, which is widely used in web development. It has little installation requirements, is cross platform and is easy use from an administrative point of view. There are thousands of user contributed modules which expand Drupal's capabilities and many JavaScript frameworks such as mootools and jQuery are inherently integrated.

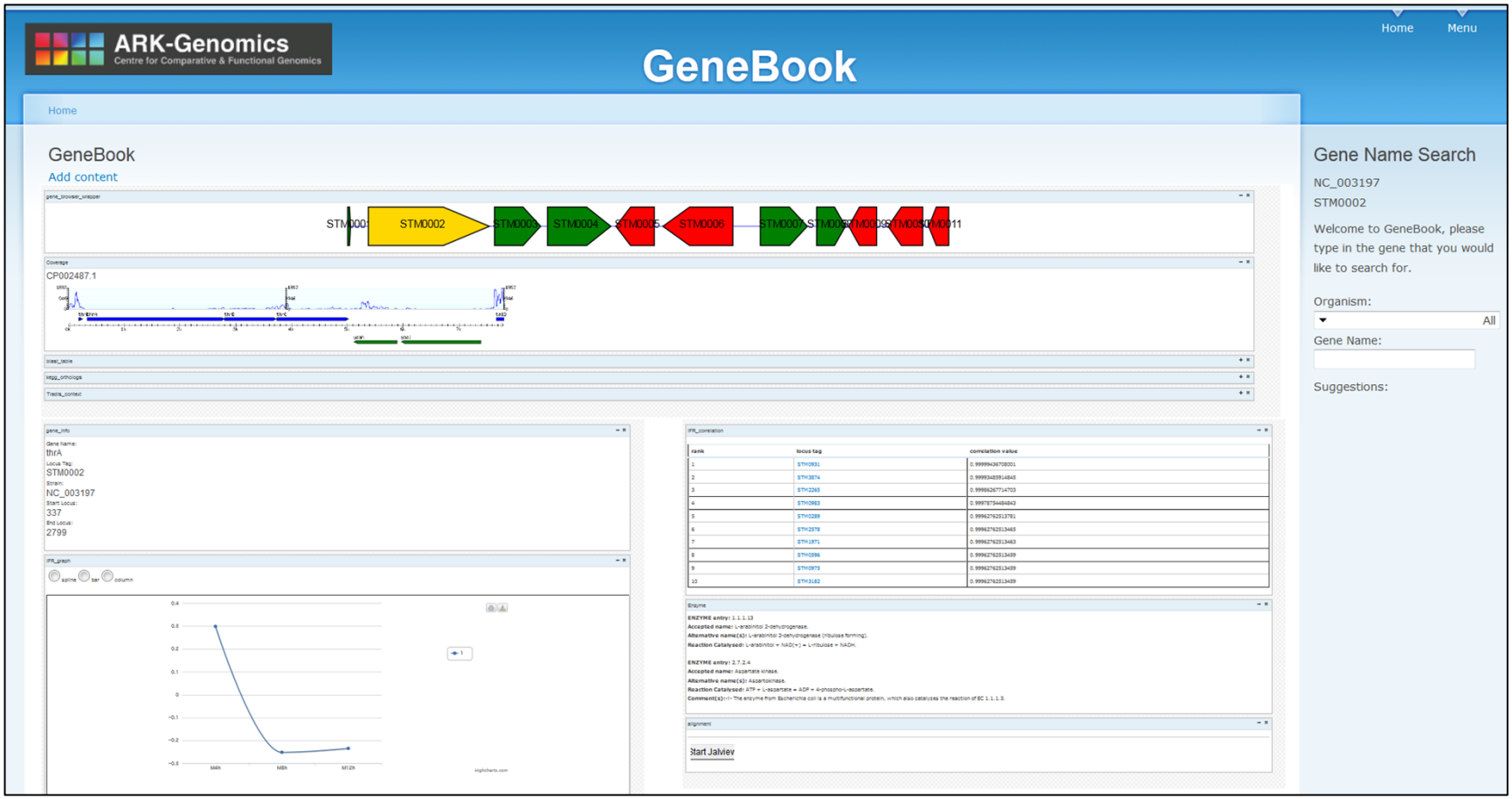


Figure 38 The interface for GeneBook. The right hand section shows the search area. The boxes in the centre are 'widgets', each widget displays fetches, parses and displays data from a remote source

4.2.1.2.2 Website Interface

Search functionality was implemented to allow the user to query the database from the web site. A drop down menu populated with the genomes from the database was integrated onto the page. Below this a text box for the user to type in gene names. The search functionality currently lets the user find a gene of their choice by first selecting a genome and then typing the gene name into the text box.

A dynamic interface is integral to the system. Web 2.0 encompasses dynamic content, user controlled design, and information sharing. Many of the websites have a desktop application appearance, rather than a static page of text.

Autopredict capability was integrated with the gene name text box. This was achieved using AJAX. With each key stroke inside the text box a JavaScript script is activated which uses the HTTP request function to query the database. The query asks the database for any genes which start with the typed letters. Those that match are returned as a hyperlinked list. The fact that the HTTP request works asynchronously means that web page does not need to be refreshed in order to see the list changing. Clicking on one of these gene names uses the HTTP request function to show information (from the database and remote data sources) about that given gene displayed in widgets.

Widgets, in the context of this project, are small boxes that can contain data from remote web services or tools for interrogating the system. The widgets are flexible, meaning the user can choose which widgets they want and where they want them to be (Figure 38).

New widgets can easily be added, as the web services interface is kept completely distinct from the web page itself. This is simply a matter of clicking to make a new block (widget) through the Drupal CMS. Each widget uses a wrapper to connect to the webservice. In the text box for the new widget the user only needs to state the name of the wrapper and any extra parameters like so:

```
<p class = 'invisible'>coverage</p>
<div class = "widget">typhim_474_sorted.bam typhim_474.fasta 7500
coverage1 </div>
```

The text in the p tag (‘coverage’) is the name of the PHP wrapper that GeneBook uses to contact to the webservice. The information in the div tag is space delimited for each parameter, this information is fed to the PHP wrapper as parameters via the POST method.

The widget boxes are made automatically using the Homebox module from Drupal [219]. Homebox widgets are draggable. Which widgets the user has chosen to activate and their location are recorded in the Drupal database, meaning that each user has a personalised space. This also ensures that the state is maintained between sessions.

To fill the widgets with data from the remote web services the JavaScript framework jQuery [220] ‘live’ method was used. This applies an event to all elements, even those which are made after the page has loaded. As the data changes with each click of a different gene, the widget elements need to be filled on-the-fly, without page refreshes.

The interface, by design, offers access to multiple different data types in a clean, unfettered environment. Behind the web browser GeneBook’s server is firing off multiple requests to many webservices and returning each set of results into a widget. This is made possible due to the increasing number of webservices available.

4.2.1.2.3 Remote Data Access

There are many tools/databases offering webservices. These allow users to use the remotely based service directly rather than via a website. This transition can be attributed to the following reasons;

- Much of the data produced at the genomic level is large and computer intensive to analyse. Using a bioinformatics tool which is located on a server elsewhere means that anyone with a reasonable internet connection can perform these analyses; the user is empowered with computing capabilities beyond that which is provided to them locally.
- The user does not have to deal with software installations, maintenance and updates which are common with desktop applications.
- There is no need for database synchronisation or update.
- Webservices are cross platform, meaning that they are not limited to one operating system/programming language.
- The user can access large datasets quickly and efficiently.
- The user just needs a web browser.
- Accessing the tool/database directly (rather than via a website) means that the user can perform fuller queries.

4.2.1.2.4 GeneBook structure

GeneBook in its entirety is divided into three main components; the database, the web interface (CMS) and the webservices. Keeping the components completely separate from each other ensures that although they can interact they are not dependent on one another. This means, for example, that the database can be accessed without the web interface.

The database and web interface simply connect using PHP's MySQL functions, the parameters for which are contained in a database parameters folder located within the GeneBook file structure.

The way that GeneBook connects webservices to the database and the web interface is more complex. Figure 39 shows a general schematic of how GeneBook's components interact. Figure 39 shows how the different aspects of GeneBook interact with one another. In order to access a given webservice a wrapper is created in GeneBook's file system. The wrapper contains the code that queries GeneBook's database, sends a request to the remote webservice and parses the returned result into a human readable format within a widget. All wrappers can be accessed directly, without being displayed in a widget in GeneBook. This independence allows users who to access the webservice and get unparsed results.

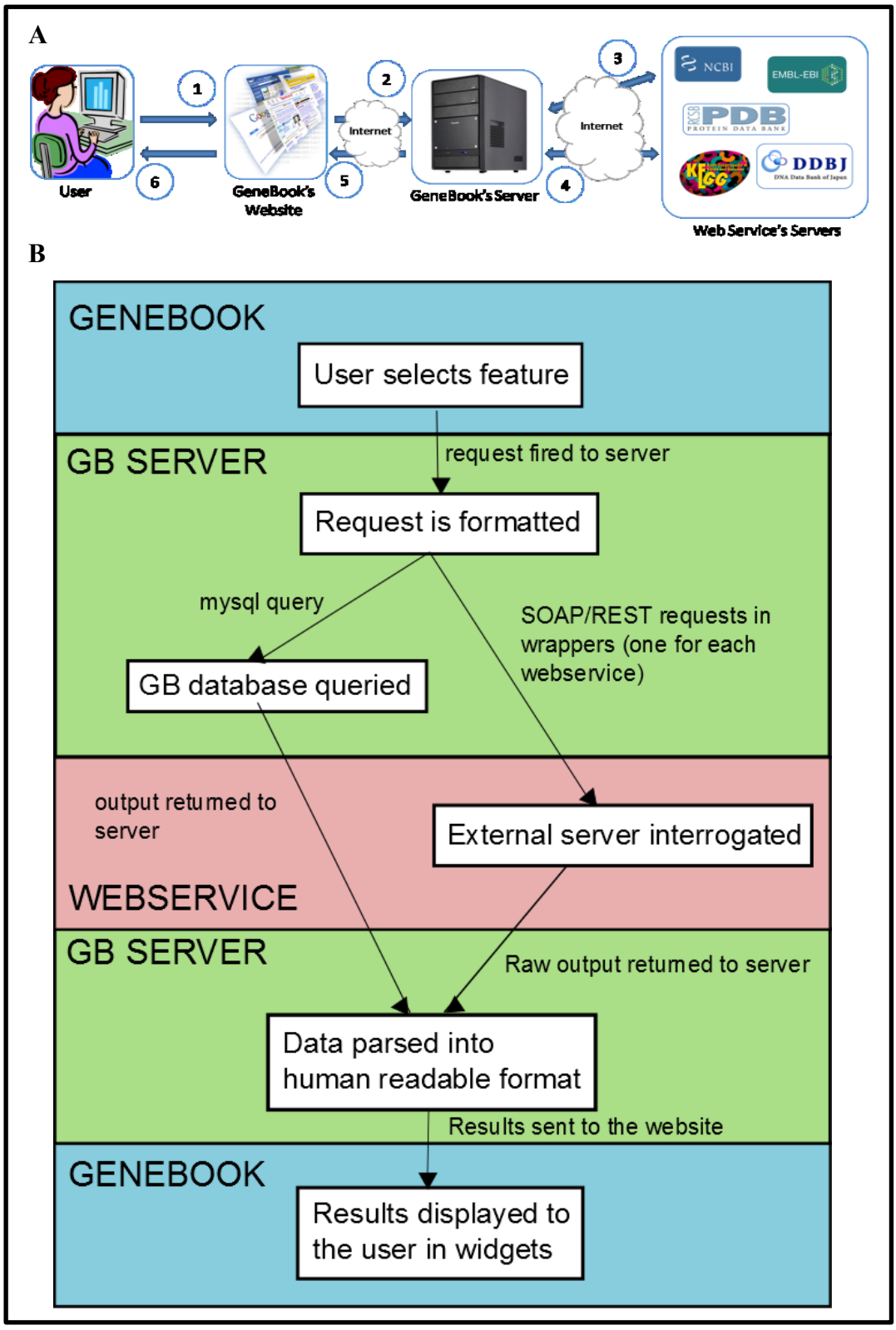


Figure 39 A) Cartoon of GeneBook interacting with webservices. User selects a feature (1). The query is sent to GeneBook's server (2), which simultaneously queries the GeneBook database (3) and fires the request to external servers. This is returned and parsed by GB server (4) and then displayed on the GeneBook website (5). B) Flow diagram showing the steps occurring when a user selects a feature in GeneBook.

Apart from the existing web services, some in house webservices were created too. That is web services that are hosted on the Roslin Institute servers. These webservices are located in the web directory but are completely independent of GeneBook (for more details about in-house webservices see 4.2.1.3.2). They are either in the CGI folder or in the Webservices folder, this is based on the whether they are programmed in Perl or PHP respectively. Although these services could be accessed by GeneBook directly they are accessed in exactly the same method as remote webservices (via wrappers). This means that that if a user wants to access a Roslin Institute webservice they can do so without having access to GeneBook or its database. Also, if GeneBook is installed locally onto a user's machine they can still access the 'in-house' webservices without having to download them or their dependencies.

To keep the GeneBook ethos of flexibility there is no hard coding of directories or files. The Webservices, CGI and Wrapper folders all contain a parameter file. This contains any directories used, meaning that if GeneBook or the in-house webservices are moved/installed locally only the parameter files need to be modified (Figure 40).

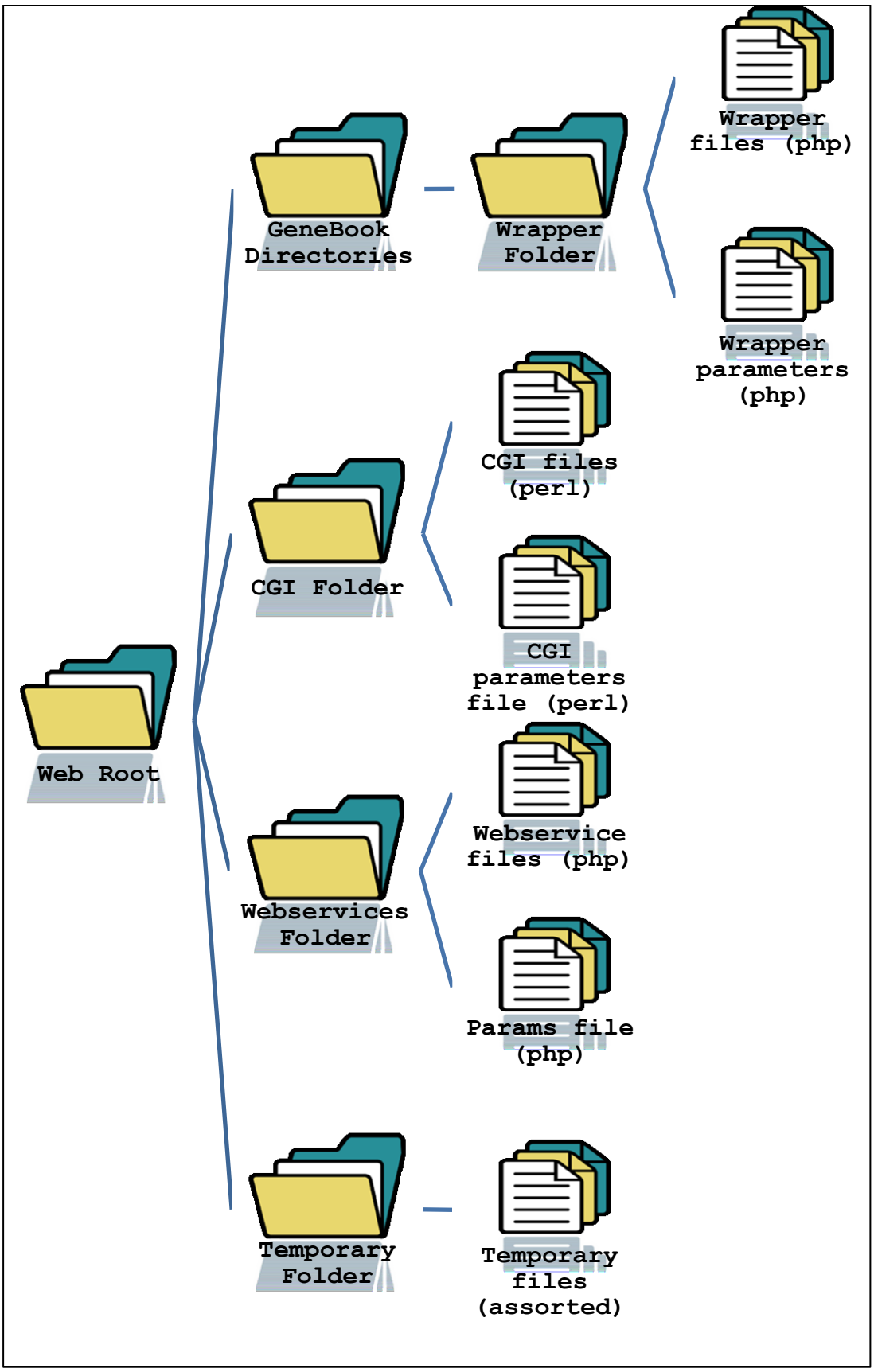


Figure 40 GeneBook file structure, showing how the in-house webservices are independent of GeneBook all widgets are made by files in the wrapper which is integral to GeneBook.

4.2.1.3 Integrating webservices

Although GeneBook accesses both remote webservices and in-house webservices as remote webservices their development and integration have been described in distinct sections. This is because the integration of true remote webservices only requires a wrapper. Conversely, the development of each in-house webservice results in both the webservice itself and a wrapper.

4.2.1.3.1 Remote webservices

These webservices are hosted externally, a wrapper was designed for each website. These send the requests and parse the results into an attractive, human readable format. In total six remote webservices were integrated, each requiring different remote access protocols and methods for displaying the data. This section describes these web services in detail including how they are accessed, what the web service returns and what this is parsed into.

ENZYME

As described in the ERD earlier (Figure 37) EC number is an entity in the database. The ENZYME database [221] offers information about all EC numbers including their name/function and the reaction they catalyse. One can navigate around the database by simply changing the URL. For example the EC number 1.1.1.1 is for alcohol dehydrogenase. To find this entry on the ENZYME database one would enter the following URL;

http://enzyme.expasy.org/EC/1.1.1.1

This data is in the form of a web page where the data is displayed in a human readable format, arranged into headings, hyperlinks and tables. This formatting means that it is hard to extract information computationally, as a parser would have to navigate around the html tags and all the non-relevant data they contain. The ENZYME database does also offer all entries as a text document with no html formatting:

http://enzyme.expasy.org/EC/1.1.1.1.txt

The ENZYME wrapper accesses the ENZYME database by parsing the EC number onto the end of the URL (followed by '.txt'). The HTTP request function submits the URL and returns the text into an array. Each line of the text file starts with a header. These headers were used by the wrapper to parse the data into a more visually pleasing format and displayed into a widget (Figure 41).

Accessing the ENZYME database using this method isn't strictly using a webservice as there is no query being submitted, this is a hardcoded URL. This wrapper demonstrates that it can be possible to access remote data even when they don't offer an API.

One of the limitations of this wrapper is that it requires an entry in the EC entity. It is possible that newly submitted genomes won't have enzyme information. Assigning orthologs to every feature could overcome this. Alternatively, software could be used which automatically predicts EC number, although these are often based on homology methods too.

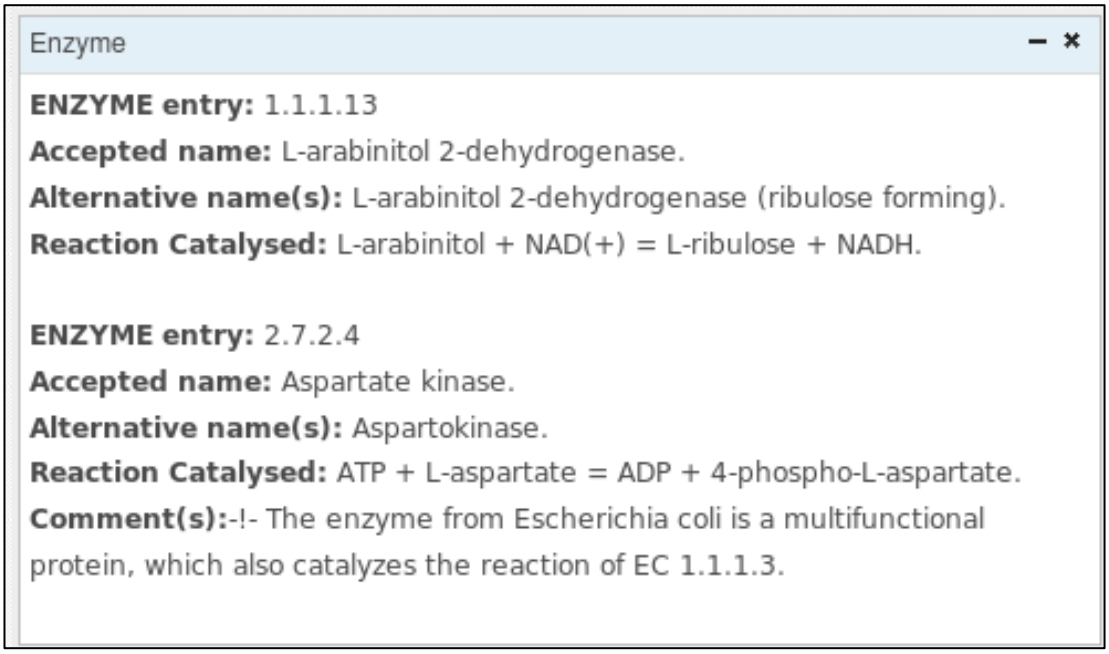


Figure 41 Widget with output from the Enzyme wrapper for STM0018, showing that there are two enzymes associated with this gene.

BLAST

Many research institutes such as the EBI have made multiple web services [127] that allow users to access their databases and tools programmatically rather than through a GUI. The tool chosen for remote access was BLAST as this is a good starting point for inferring gene function, the EBI's API use SOAP.

The wrapper creates a SOAP client using the EBI's NCBIBLAST WSDL file.

http://www.ebi.ac.uk/Tools/services/soap/ncbiBLAST?wsdl

The sequence is obtained from GeneBook's database. The parameters for BLAST (sequence, program, database and sequence type) are then defined in an array. The SOAP client submits this array as a query to the EBI API. The wrapper uses a loop to poll the status of the job submitted, once the results are ready the output is saved as both a visual SVG (Figure 42) and an XML table. These outputs are parsed into separate widgets, visual BLAST and a table showing the results.

Although SVG files appear as images they are actually in an XML format which is rendered by the browser. The visual SVG files are parsed so that each BLAST bar has an embedded hyperlink to the protein it hits in Uniprot [222]. This is then displayed as a widget with clickable proteins.

The XML file is parsed into a table displaying the hit (with a hyperlink) and the scores. This is displayed in a separate widget, offering the users two ways of visualising the BLAST outputs (Figure 43).

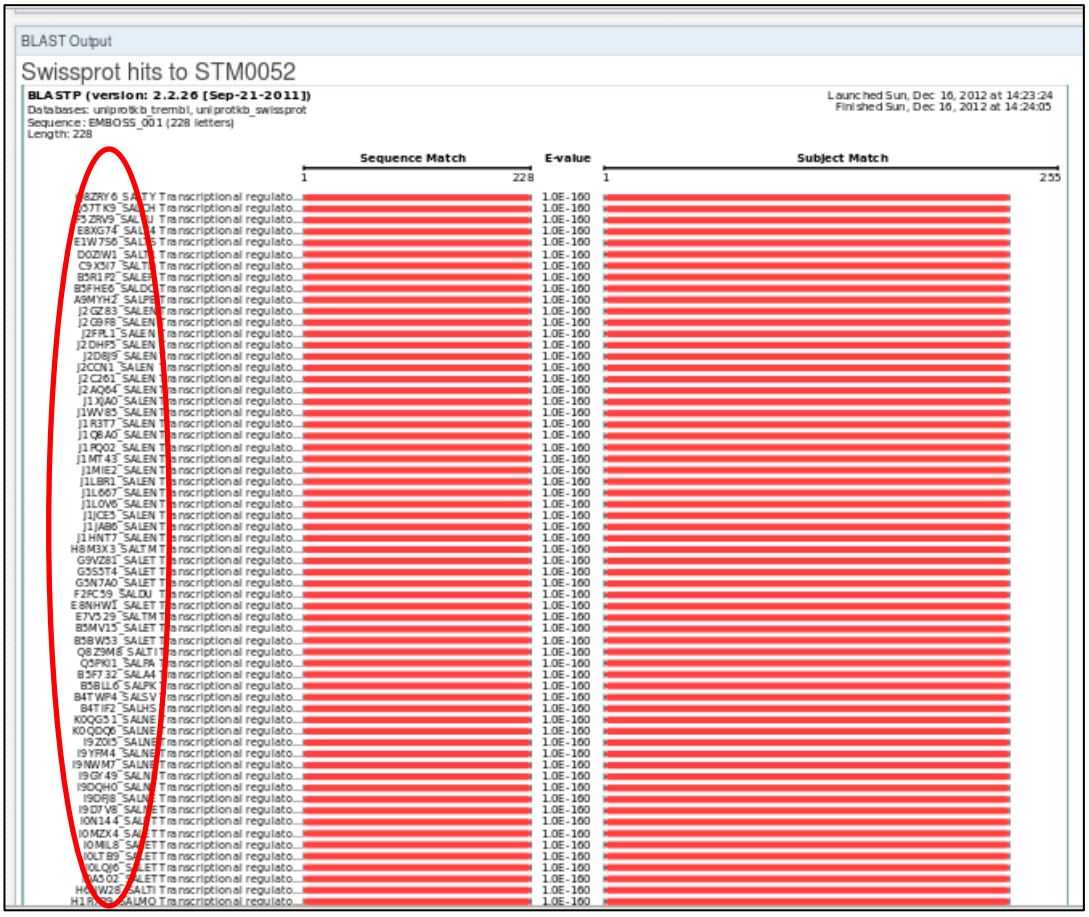


Figure 42 BLAST SVG output for STM0052 in a widget, the Uniprot IDs (circled in RED) have an embedded hyperlink allowing users to click on these for more information about that protein.

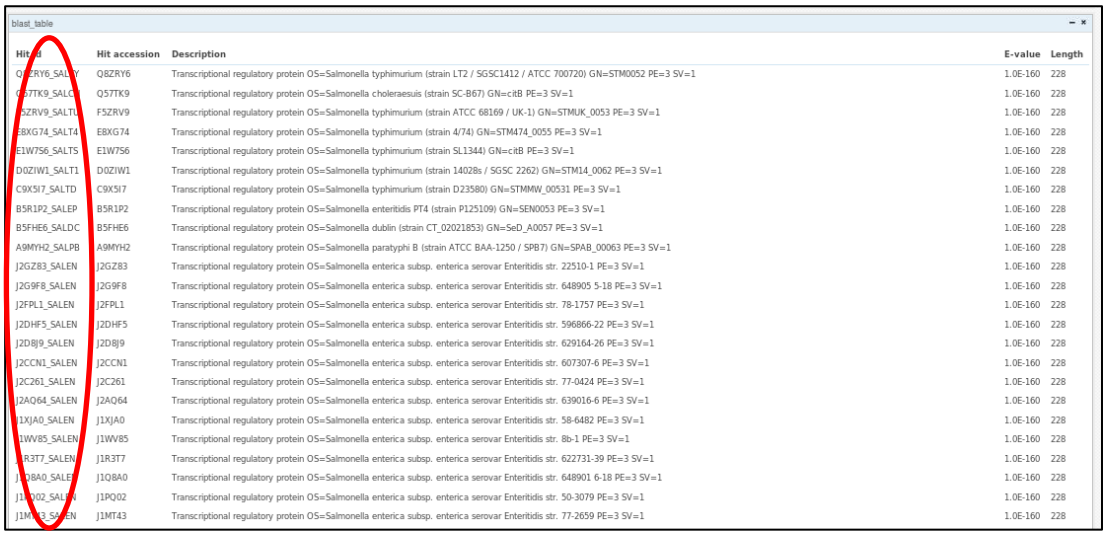


Figure 43 The BLAST output for STM0052 in table format, showing the e-value, protein description and length of the match. The Uniprot ID is circled in red, these are hyperlinked to UniprotKB, allowing the user to easily get more details of the protein of interest from within GeneBook.

GEO

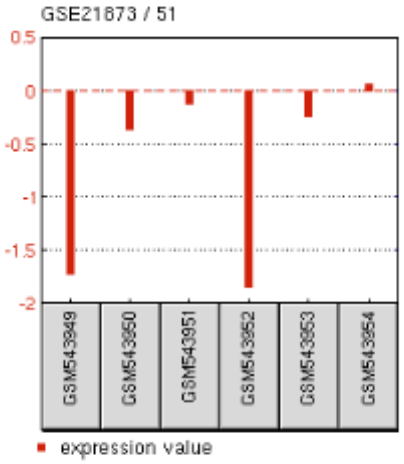
The Gene Expression Omnibus (GEO) [223] is a microarray data repository. Although it doesn't have a true RESTful API, GEO can be queried using the URL. Microarray data is complex, in each experiment the data is organised into series, platform and sample.

In terms of writing a wrapper for this there are two limiting factors. Firstly, GeneBook sends queries based on the locus tag or gene name. This information is stored in the GSE data, the data points themselves are in a different file, GPL, and are organised by a unique identifier (locus tag wouldn't be unique because there can be more than one spot for each gene). The second problem is visualising the data, GEO does have the GEO2R tool that displays the data points for a given GSE ID dynamically in a graphical format, however, this is not designed to be accessed programmatically.

To overcome the first issue the wrapper firstly gets the XML of the GSE data and finds the rows associated with the locus tag of interest. The rows have to be searched because the gene identifiers (in this case locus tag) are not consistently in the same column or under the same header. The IDs were parsed more easily because they are always the first column of every row. The next issue was tackled by looking at the

HTML code for GEO2R it became apparent that the graphs it makes are generated by a CGI script. The URL for the CGI was incorporated into an iframe tag within the wrapper and the IDs discovered from the previous step were parsed into the URL as parameters (Figure 44).

Profile GSE21873 / 51
Title Salmonella Typhimurium LT2: Control versus naringenin
Organism Salmonella enterica subsp. enterica serovar Typhimurium str. LT2



Graph caption help

Sample	Title	Value
GSM543949	Naringenin vs Control Replicate 1	-1.7412
GSM543950	Naringenin vs Control Replicate 2	-0.379683
GSM543951	Naringenin vs Control Replicate 3	-0.143946
GSM543952	Naringenin vs Control Replicate 4	-1.86416
GSM543953	Naringenin vs Control Replicate 5	-0.261776
GSM543954	Naringenin vs Control Replicate 6	0.0690145

Figure 44 GeneBook widget displaying output from the GEO wrapper. This information has been parsed directly from GSE and GPL files in GEO.

KEGG

KEGG also provides API access to their data. Above and beyond the identical genomic information that is held many similar repositories, KEGG also offers pathway data.

The wrapper submits the locus tag to the KEGG server which returns the names of the pathways associated with this gene. In order to make the service more informative another request is sent which gets the descriptions of the pathways. This means that a user can see the pathway ID and its description.

The pathway ID is sent to the KEGG API one last time, this is to get the image and colour the specific gene select. This link is embedded in a JavaScript fancybox which gets the image on the fly and displays it dynamically (Figure 45 and Figure 46).

The functionality of KEGG's SOAP API is somewhat restrictive. The programmer can access a lot of information but in order to construct complex queries they have to send multiple requests to the server, which limits the speed of the widget.

KEGG have now migrated to a RESTful API, in theory this should offer programmers the ability to construct their own queries. In terms of GeneBook the widget would be faster and would only need to send one request.

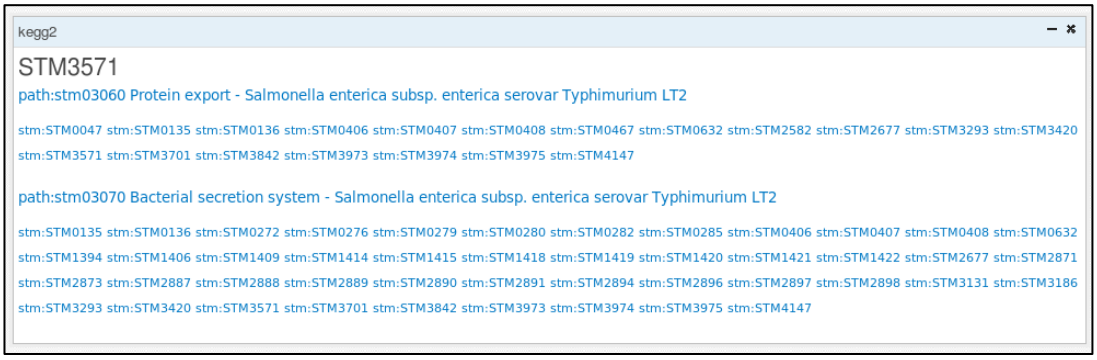


Figure 45 Output of the KEGG wrapper in a widget. The output shows the pathways for STM3571 and the other genes from this genome that belong to the same pathways. Clicking on the pathway heading open up the pathway map dynamically in GeneBook (Figure 46). Clicking on another gene open a new page in GeneBook for this gene.

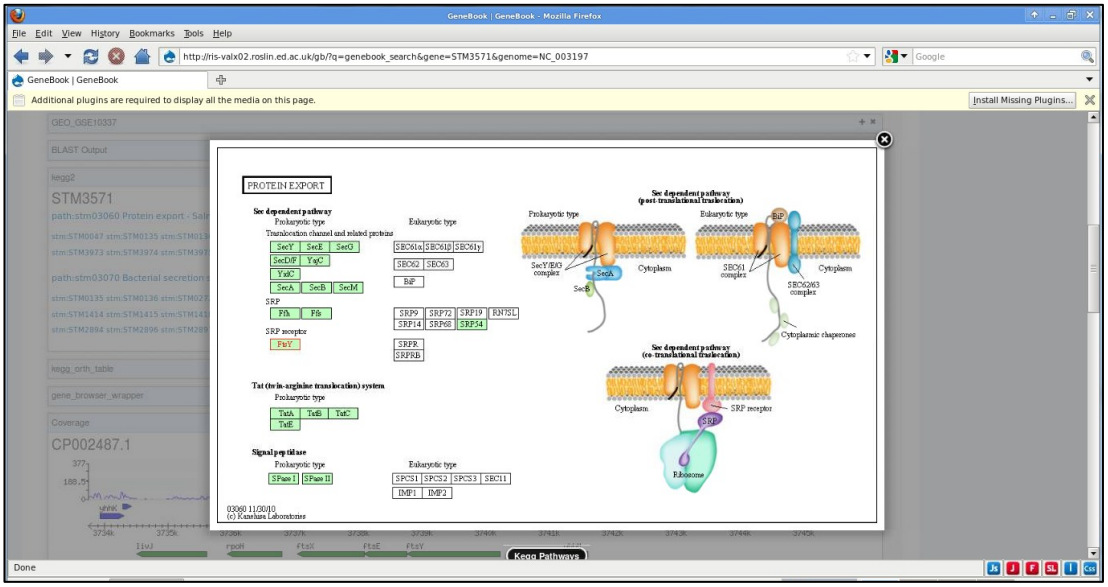


Figure 46 A KEGG pathways map for STM3571 overlaying GeneBook. This is brought to the user on the fly. The gene of interest, STM3571 is highlighted in red so that users can see what role their gene plays in the pathway.

ClustalW

ClustalW is also part of the EBI SOAP suite and was accessed in a similar way to the BLAST wrapper. ClustalW requires an array of sequences for the alignment. The wrapper currently sends a request to KEGG for orthology, this does limit speed but also keeps with the GeneBook paradigm of holding minimal information locally. If in the future all orthologs are calculated then this step can be taken out of the wrapper.

The two outputs are an alignment (Clustalw) file and a dendrogram file (dnd). The user is then able to view these in a Jalview applet [224, 225]. This is a java alignment application which can be run through the browser. This allows users to view the alignment and dendrogram on the fly in fully interactive windows (Figure 47).

One current limitation of this tool is that it relies on KEGG orthologs which do not include pseudogenes. If the gene of interest is a pseudogene this will not be identified as having any orthologs and if any genomes have an 'ortholog' which is a pseudogene this will also be omitted. With this in mind a buffered Clustalw widget was developed.

ClustalW with buffers

This widget works in the same way as the previous widget but the user can define how much sequence to include in the alignment up and downstream of the feature of interest. This means that one can look at a pseudogene by selecting the feature next to it and including a buffer large enough to encompass the pseudogene.

The user can also define whether they want the amino acid sequence or DNA sequence.

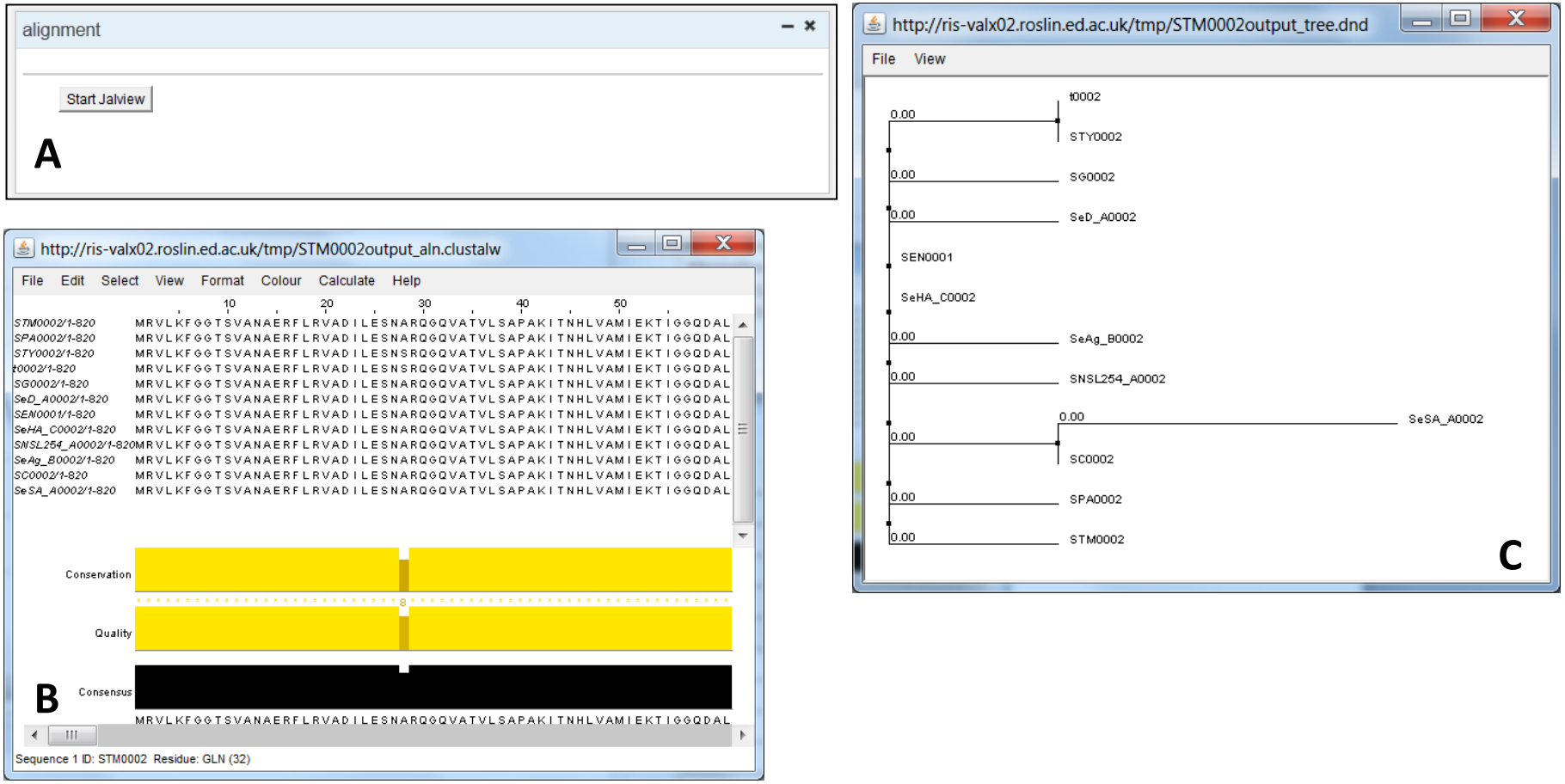


Figure 47 Widget containing the ClustalW applet, Jalview is launched by clicking on the 'Start Jalview' button (A). This opens two windows, the alignment and the dendrogram, displaying the data calculated from ClustalW.

E-utilities – Conserved Domain Database

The Entrez Programming Utilities (E-utilities) offer an interface to the Entrez system which is comprised of 38 biological databases [226]. The eutils CDD (Conserved Domains Database) widget is a proof of concept, demonstrating how any of the Entrez databases can be accessed using E-utilities. The Conserved Domains Database consists of in-house NCBI-curated domains and several external domain resources like Pfam and SMART [227].

The CDD is accessed using E-utilities' RESTful API. The feature's locus tag is submitted to the API, which returns a gene ID, this is submitted to the CDD and a list of matching domain IDs are returned, these are then submitted to the CDD to get the description for each match. This is returned in XML format which is parsed into a web list. The list is linked to an accordion JavaScript, meaning that the results are returned to the user as a list of domains, to get more information the user clicks on a domain and the list expands to show the description for this domain (Figure 48 and Figure 49).

The limitation of this widget is that it relies on the gene ID being available in the database. If the sequence could be submitted this would be make the widget extendible to genomes that aren't publically available. Alternatively, an ortholog could be submitted when the Gene ID isn't available would also overcome this problem.

The next step in the development of this widget would be to include some kind of visualisation. This could be done using the Perl Bio::Graphics package [228].

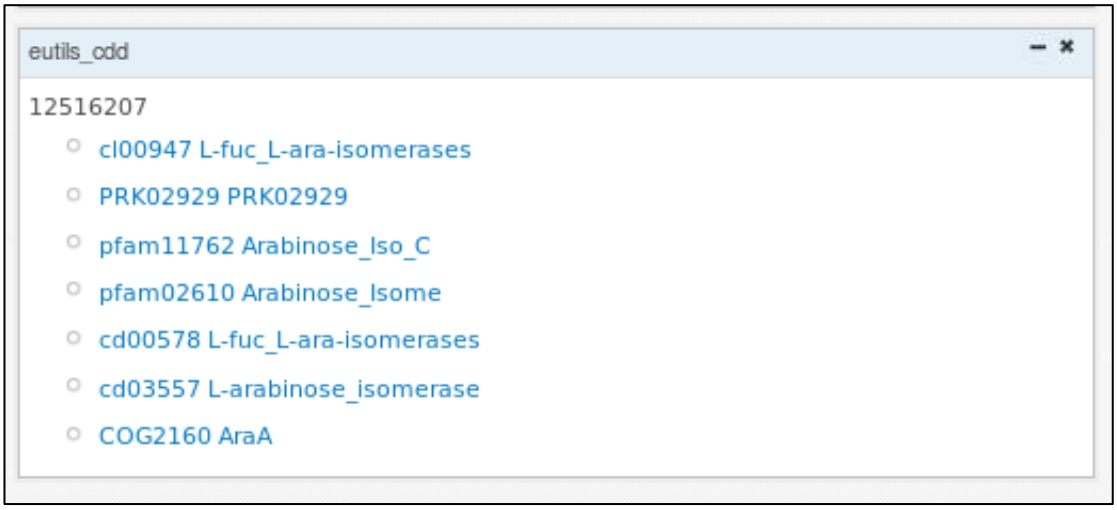


Figure 48 Output of the E-utilities CDD wrapper displayed in a widget. This list of domains is expandable/collapsible (Figure 49) by clicking on the domain of interest.

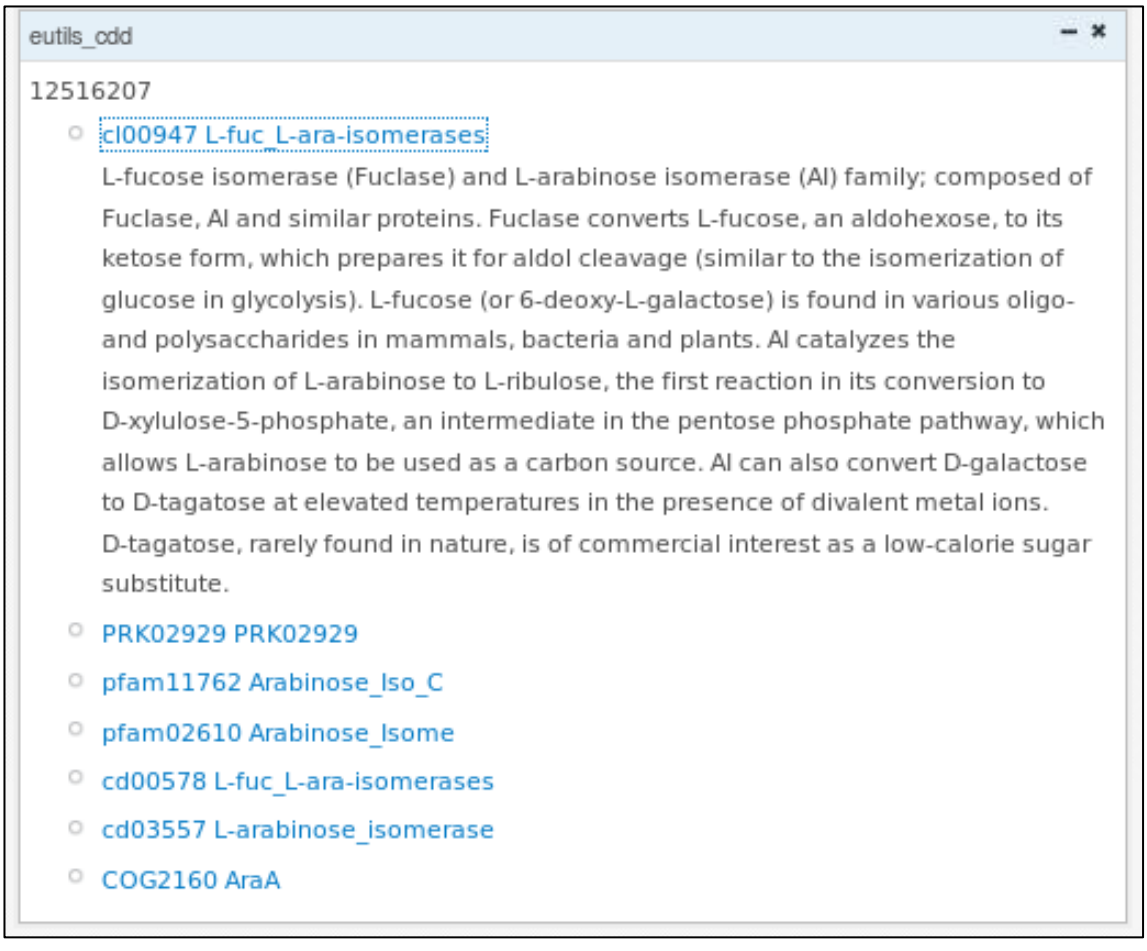


Figure 49 Widget with an expanded domain in the E-utilities list, this can be collapsed again by clicking on the heading.

4.2.1.3.2 In-house webservices

In-house webservices are distinct from remote webservices in that they are located on the same server as GeneBook and they have been created as part of this project. The development of in-house webservices required not only a wrapper but the design and implementation of the webservice itself. For this reason these webservices are explained in a section separate from the standard remote webservices. As previously explained in section 4.2.1.2.3 they are treated and accessed the same way, that is, via a wrapper even though the in-house could accessed directly. This means that if in the future GeneBook is migrated to a different server but the in-house webservices are not GeneBook will still be able to access them without having to make any changes to GeneBook or its wrappers.

Some of the webservices have been written use Perl CGI and some have been written in PHP, these webservices are kept in different folders and both have accompanying parameter files, meaning that all parameters are completely independent of the webservice scripts, this makes moving and updating GeneBook easier (Figure 40).

4.2.1.3.2.1 Basic webservices

The first webservices for this project were developed as tools for visualising the features in the context of their surrounding features and in the context of other bacterial strains. There are already methods for graphically displaying sequences [228, 229] but they don't offer a method of easily making and displaying images on the fly into a web browser. One of the aims of the project is to use web 2.0 technology, in this respect this would include interactive images with clickable regions, and again the currently available methods don't offer this without hacking.

Genebrowser

This webservice produces a hyperlinked SVG file of the selected feature and its surrounding features. A query is submitted to the GeneBook database, which returns the feature location and the features within 10,000bp up and downstream of the feature. The results are parsed into a .dot file (dot-bracket notation) and submitted to Graphviz, a graph visualisation tool [230]. The output (Figure 50) is an SVG image

with clickable features. When the user hovers over a feature the annotation description is displayed.

This widget could have been made using Bio::Graphics or another similar tool such as GenomeGraphs or ggbio [228, 229, 231]. The reason these packages were not used is primarily because it is harder to incorporate hyperlinks but also the development of this widget shows the extendibility of GeneBook, any kind of visualisation tool can be integrated.

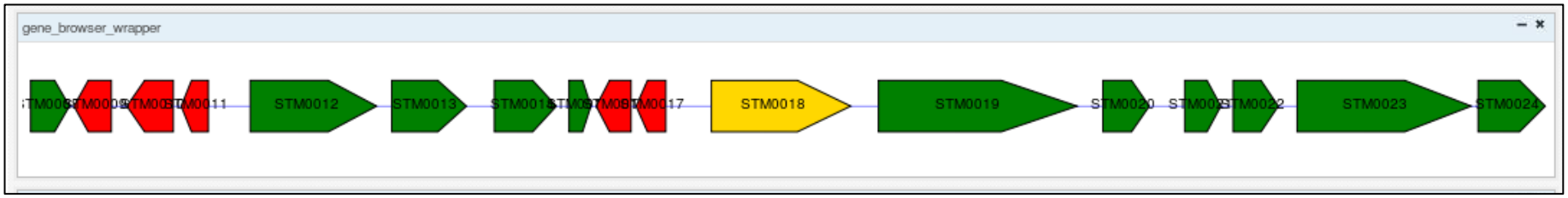


Figure 50 Genebrowser output displayed in a widget. The selected feature is highlighted in yellow, green show features on the forward strand and red on the reverse strand. All features are clickable, opening an instance of that feature in a new tab in GeneBook. Hovering over the feature will return its locus tag, and gene name.

Visualising KEGG Ortholog

The KEGG database assigns orthologs for the majority of genes residing within it, these are referred to as KO (KEGG Orthologs). This widget sends a request to the KEGG API for orthologs from the other genomes in GeneBook. The ortholog locus_tags are returned. Each ortholog is made into a clickable SVG, using the same method as the GeneBrowser widget. These are then displayed alongside one another, showing the features in the context of one another and their own genome (Figure 51).

Figure 51 output from the KO_genebrowser webservice displayed in a widget. The top sequence is the genome we are looking at (in bold top left). Yellow and blue features show the orthologs for the feature organised by host generalist and host restricted respectively. Green features are on the forward strand and red features are on the reverse. Pink features are pseudogenes.

Sequence

Sometimes it is useful to have the sequence of a region of interest. This can be used to submit to other tools or for sequence searches in the ClustalW with buffer widget. This widget queries the GeneBook database to get the feature location and its orientation. Using Bioperl, the results from the database query are used to get the sequence from a FASTA file of the whole genome sequence. The user can state whether they want an amino acid sequence or DNA sequence, how much sequence up and downstream and the orientation of the sequence (the default for these is DNA, 5000bp and the natural orientation of the feature).

4.2.1.3.2.2 Generic quantitative webservices

There has also been a surge in other next generation techniques such as RNA-seq, incorporating experimental methods gives a better indication of a protein's role and whether it is functional. These annotations would be more accurate because they are based on actual experiment data rather than homology. Currently genomes can include evidence tags stating how the annotation was assigned, however, they are often omitted from the process. Including evidence qualifiers gives the user an idea of the reliability of the reference genome. The concept of assigning a level of quality to annotation is not novel, but is seldom used [159, 168].

This section explains the in-house webservices that have been developed to handle a user's quantitative data. The premise is that they can submit a basic tab-delimited file which can interpret and visualise their data. This can be any kind of data as long as it is quantitative such as time-course, microarray or mutagenesis data.

There is no prior knowledge of computing required, the file could be made in a common program like Microsoft Excel. These widgets work by the user changing the headers of their data to conform to the webservice's protocol.

Graph

The purpose of this widget is to plot generic quantitative data, this is achieved by using the JavaScript library HighCharts [232]. This library is capable of making figures on the fly, no installation is required and it allows the user to download the output.

The steps for making the data suitable for visualisation are as follows:

- The data must be saved in a tab delimited file (this is an option in Microsoft Excel).
- There must be a column which has the locus tags of the features which were analysed in the experiment.
- The columns which hold the numerical results must be labelled data_XXX where XXX is the label for that piece of data.
- If there are p-values for the data points, they are labelled as pval_XXX where XXX is the label for that piece of data

The rows which correspond to the feature of interest are extracted, parsed according to their headings into a HighCharts compatible format. The image is then displayed in the widget. The user can see the p-value and whether it falls into the range of significance. Because the graph is made on the fly the user can chose which data points they see and what kind of graph is best for visualisation. Figure 52, Figure 53 and Figure 54 show the visualisation of different types of data using the same generic graph webservice. Our in-house widget can display any type of quantitative data, is very fast and lightweight and allows users to download the output.

One possible limitation with this method is that the experiment has features labelled as ORF1, ORF2 etc. rather than their true name. The question is how does the user know what ORF1 really is, presumably they do a sequence alignment or integrate their results with the true locus tags. It is not unreasonable to expect users to include the locus tag somewhere in their results.

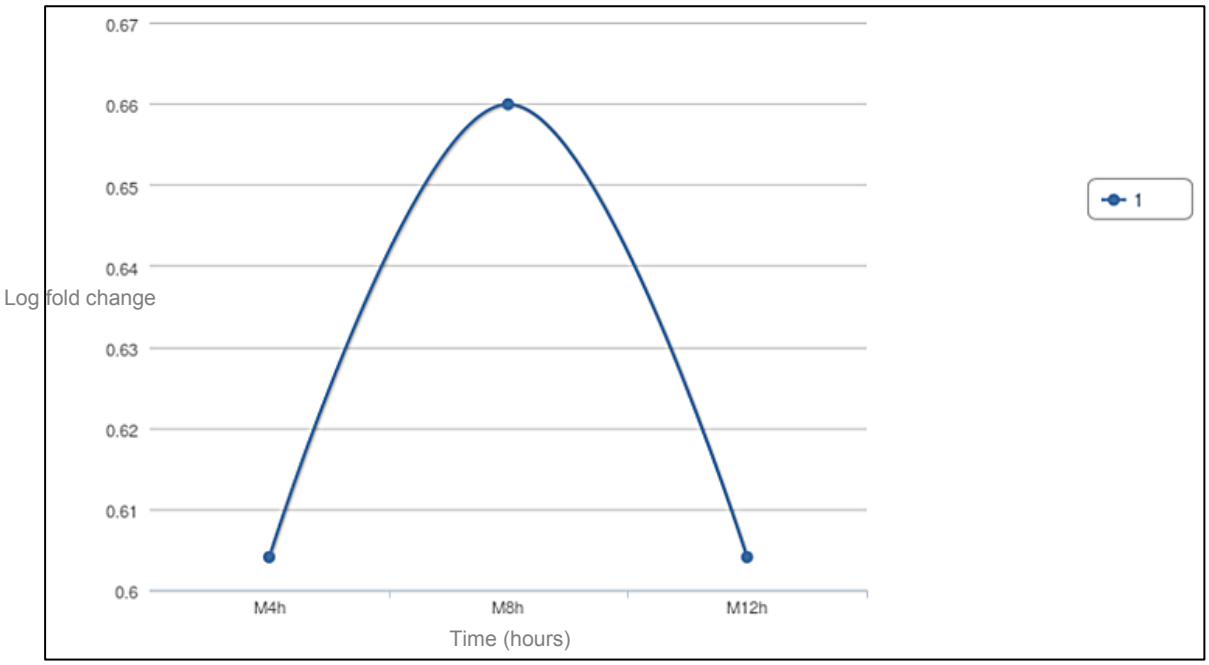


Figure 52 Time course data displayed using the generic quantitative data widget . This is in a widget created on the fly by GeneBook, nothing is pre calculated.

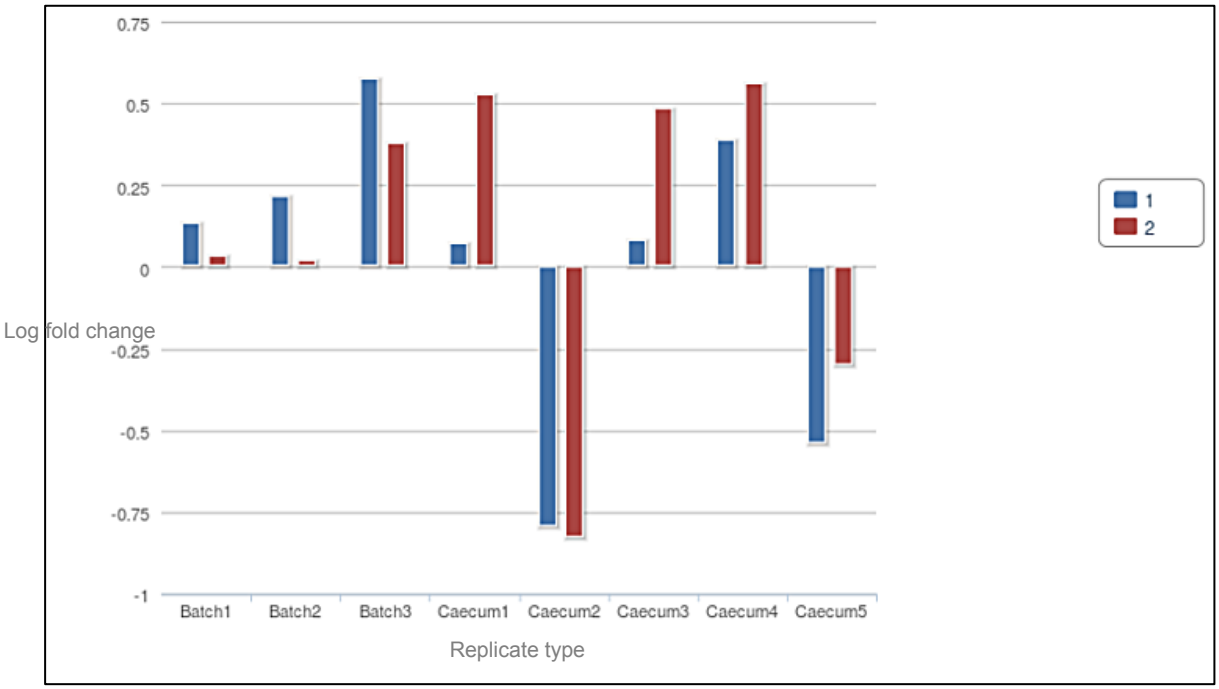


Figure 53 Widget displaying private microarray data as a column graph showing replicates for batch and caecum. This is in a widget created on the fly by GeneBook, nothing is pre calculated.

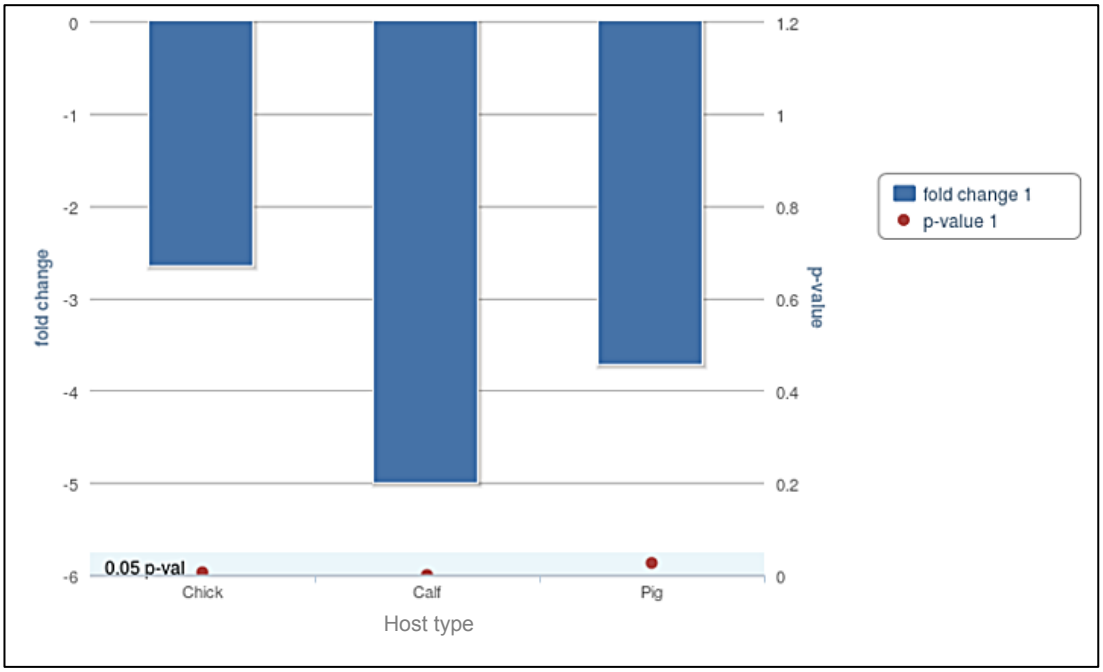


Figure 54 Graph made in GeneBook using the quantitative graph widget showing data with p-values. This is in a widget created on the fly by GeneBook, nothing is pre calculated.

In terms of improving this webservice the first step would be to allow for p-values that are for the row of data of interest rather than just the data points. For example some microarray experiments are designed this way. This could be visualised by showing the p-value for the entire graph and highlighting whether it is significant. Allowing the user to easily see if the experiment is of interest for this feature.

Another point of extendibility would be to find the data of interest based on sequence location. This could be used mutagenesis data which may have point mutations but no locus tag.

Location/sequence context

Another way of visualising data is in the context of the sequence. For example counts for sequencing data or point mutations in mutagenesis data. Two widgets were developed for this to demonstrate the different ways that quantitative data can be visualised. One was designed for generic quantitative data and the other was designed to visualise NGS data.

The first webservice is for generic quantitative data to be visualised in the context of its sequence (Figure 55). The rules for the data are identical to the previous webservice's, except a position_XXX heading is required too. The webservice gets the gene location (start and end) from the GeneBook database. These numbers are used as the range along the x-axis. Each position data point is used as the point for a scatter data point on the graph with the data_XXX being the number for the y-axis. If the user wants the orientation of the position to be visualised they add an orientation_XXX column which holds either '+' or '-' for forward and reverse orientation respectively.

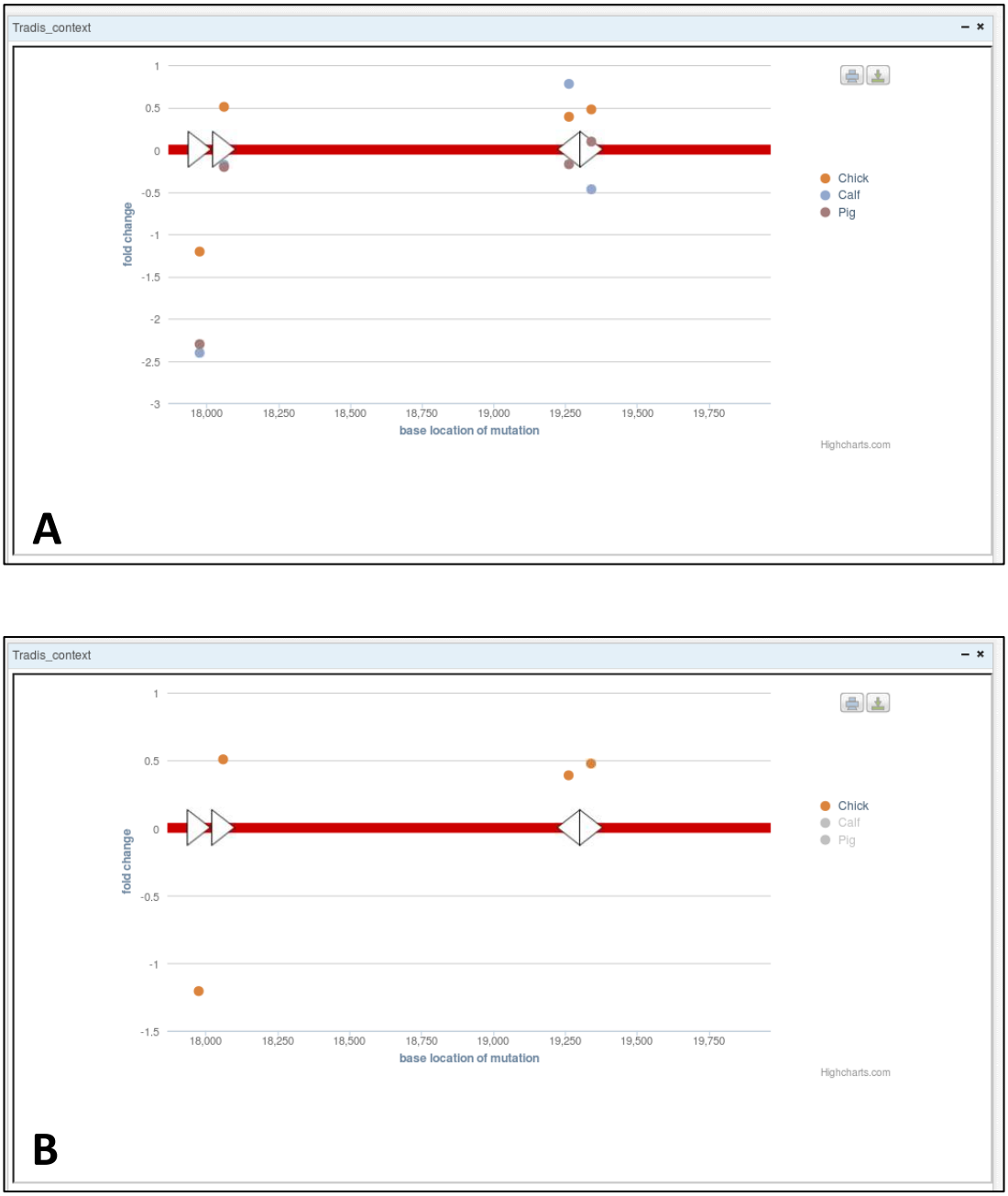


Figure 55 Gene context graph created by GeneBook showing data which has location information. In this example this is mutagenesis data where the white triangles show the point (and direction of mutation), the coloured points shows the fold change for different hosts (A). B shows that the user can select what data they want to see, this is achieved by clicking on the legend to the right, in this case the user has selected to only see chick data. Both of these images are in a widget created on the fly by GeneBook, nothing is pre calculated.

The other gene context quantitative data webservice is for NGS data. In theory if the user has made a count file for their NGS data this could be displayed in the former webservice. However, it isn't reasonable to expect users with little knowledge of handling NGS to convert their file to counts. A count file is the number of reads that overlap each point in the sequence. The count files could be precalculated by GeneBook but this veers away from its dynamic, lightweight paradigm. Calculating a count file for the entire NGS output isn't quick, which is why this webservice was developed separately from the generic quantitative data. It takes the NGS BAM file and just opens up the region of interest using an index, this is much more efficient.

This webservice gets the feature of interest and the features up and downstream within the defined range. These features are made into a 'track' of the Bio::Graphics package, they are coloured according to their orientation and if they are pseudogenes. Bioperl has a package that deals with NGS BAM files, Bio::DB::SAM this is used to open the region of interest, it can handle BAM files stored locally or those on an ftp site. This package can automatically calculate counts for the region. The counts are then made into another track as a wiggleplot (Figure 56). This image is then saved into the temporary folder, allowing users to download the image if they require it.

The fact that this webservice can handle ftp files reduces unnecessary file duplication and saves time in terms of uploading very large files.

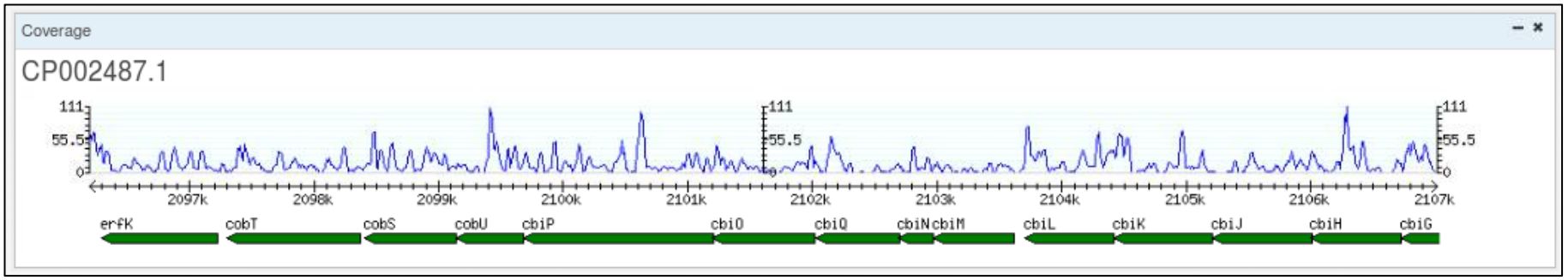


Figure 56 Coverage plot made by GeneBook and displayed in the GeneBook coverage widget. This has used the genome annotation and RNA-Seq results.

Correlation

This widget is based on Spearman's rank correlation coefficient. This is a way of comparing two sets of non-parametric data for their dependence. This webservice takes the data for the feature of interest makes it into an array and performs Spearman's correlation against every other locus tag's array of data. The results are then ordered by score.

This webservice uses the generic data rules outlined earlier, meaning it requires clearly defined heading so that it can get the locus tags and data points.

Basically, this webservice can give a list of features that show a similar pattern in the data. For example, this could be used with time course data identifying features which show a similar pattern of change over time. The variable analysed is the feature or gene in question, but in theory, as this widget is a generic data widget, it can be used on any tab-delimited data set that has multiple data points.

The data is outputted into a table showing the features in order of correlation. The table consists of hyperlinks to the corresponding features and the correlation score. Currently the significance isn't calculated for these results but a p-value could easily be implemented into the output using a statistical test like a t-test.

The resulting feature list would give the user an idea of what features are behaving in a similar way in a given experiment. If they have data for multiple experiments, it would be possible to see the results in the context of one another and if some features are appearing repeatedly across all experiments. From this the user can infer that these features are linked in some way, the feature allows users to form hypotheses from multiple data sets.

Integration of NGS data

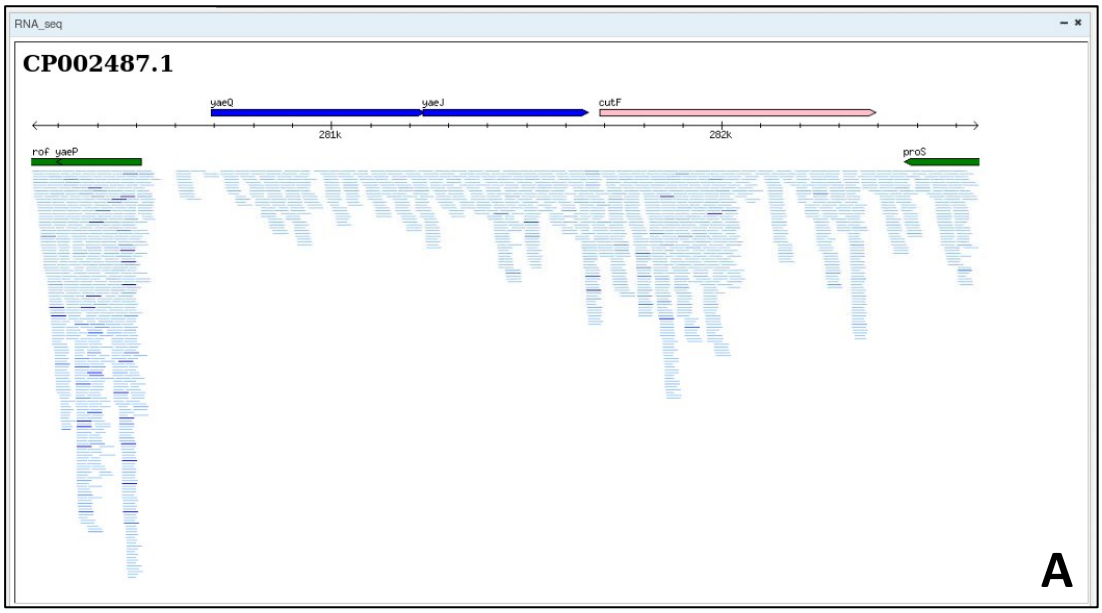
As previously described NGS datasets are large and in a standardised format. With that in mind it was logical to make a webservice and widget separate from the

generic quantitative webservices. It is capable of accessing data uploaded to GeneBook and files from remote ftp.

This webservice uses the Perl Bio::Graphics package and the Bioperl module Bio::DB::SAM, which handles SAM and BAM files. The locus tag of interest, file name of the BAM file and buffer are the main parameters. The sequence context for the region defined by the buffers and locus tag is made using Bio::Graphics. The same region is opened for the BAM file using Bio::DB::SAM. All the reads for this region are made into objects in bioperl. These are laid onto a Bio::Graphics track and displayed with their paired reads.

During the course of developing this webservice became apparent that visualising stacked paired reads can produce very large diagrams in areas of high coverage. These diagrams were too big to look at in the context of other data. With that in mind a method of condensing the data was used. That was to overlay identical reads rather than stack them. The colour of the reads is a range from light to very dark blue. A light coloured read shows that there are not many overlaid reads, on the other hand, a dark coloured read shows that there are lots of identical reads that have been aligned to exactly the same location (Figure 57A). When the user hovers over a read with their cursor the paired reads are highlighted in red (Figure 57B). This is achieved by making an html image map which dynamically displays the matching paired reads using jQuery.

The fact that this webservice uses a condensed method for displaying the data means that it is fast even over regions of high coverage. It allows users to look at the NGS data in the context of other data, without having to open a specialised NGS browser such as Tablet or IGV [55, 56].



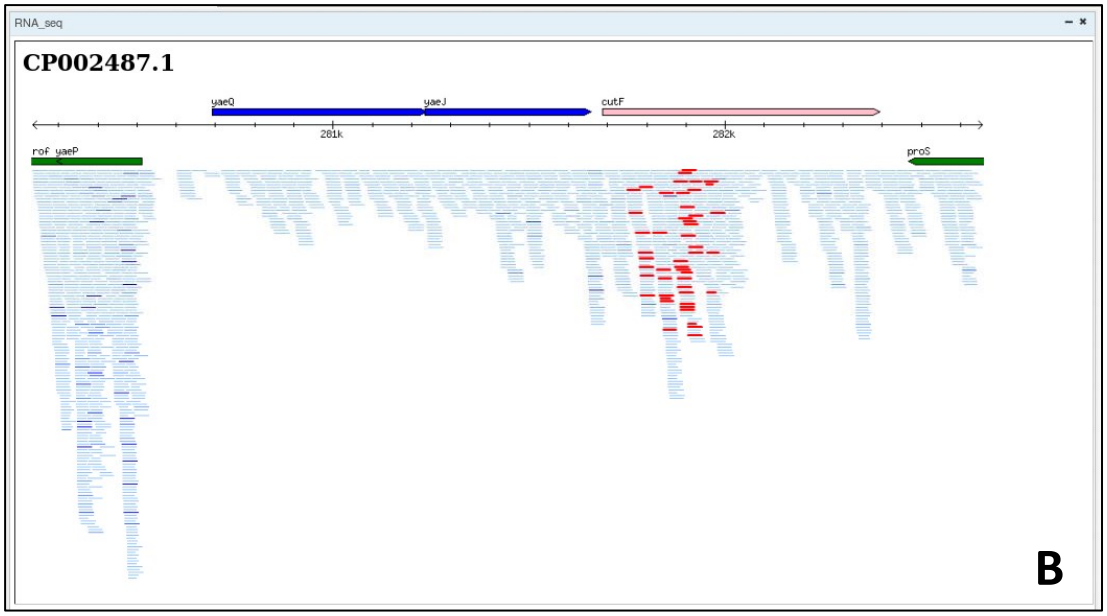


Figure 57 NGS pileup made by GeneBook displayed in a GeneBook widget. Image A shows pileup of NGS reads. The darkness of blue indicates the number of reads that occupy exact the same location in the alignment. Overlaying identical reads makes the image easier to view without scrolling. Image B shows highlighting when hovering over a read, The red reads correspond to all the paired reads for the read location that is being hovered over by the cursor.

4.3 Using GeneBook to identify annotation inconsistencies and genes associated with host pathogenicity and pathogenicity

Current methods of bacterial annotation produce errors and annotation discrepancies (Section 1.4.2). GeneBook can integrate remote data sources with the original annotation to display the most recent information available. It is capable of pulling together information from an unlimited number of resources, highlighting possible errors, inconsistencies and obsolete annotations, offering the most up to date information available. Section 4.3.2 uses real examples of the annotation problems in earlier sections to demonstrate how GeneBook overcomes some of these annotation problems using public and private data.

After performing the analysis of an experiment researchers often have a list of genes of interest. At this point they only have a locus tag or gene name. In order to find the relevance of these genes the researcher will go to multiple websites/tools to elucidate the genes' function and behaviour and try to draw conclusions in the light of their results. Further to this it is not unusual for the user to want look at the results of previous experiments in the context of their current results. Seeing all of this information can involve running multiple programs, opening many tabs in the browser and using multiple spreadsheets. Once this is repeated for each gene, it is a laborious process.

As previously described in section 4.2.1.2 GeneBook is capable of integrating local and publically available data. The examples below use data, some of which is publically available but has been stored locally to demonstrate how one would integrate their local/private data into GeneBook. Section 4.3.2 takes some of the significant genes of interest from the results of locally stored experimental data for *Salmonella* Typhimurium LT2 to show how GeneBook can be used to augment the findings with public data. The data are as follows:

- TraDIS – this has been explained in section 3.2.3.5. Essentially it shows the fold change in attenuation of Salmonella mutants across different hosts.

- Macrophage data – Microarray data showing complete transcriptional profile of S typhimurium genes at 4, 8 and 12 hours post-infection of murine macrophages [133]
- Caecal data – Microarray data comparing the expression of all S typhimurium genes in bacteria harvested from the chicken caecal mucosa with bacteria growing at log-growth phase in broth culture [94]

All of the examples shown in this chapter are from widgets within GeneBook, for ease of viewing in this document the individual widget results are shown as separate figures. In GeneBook all the widgets would be in one browser page for each example, Figure 58 shows an overview of how these results would be displayed.

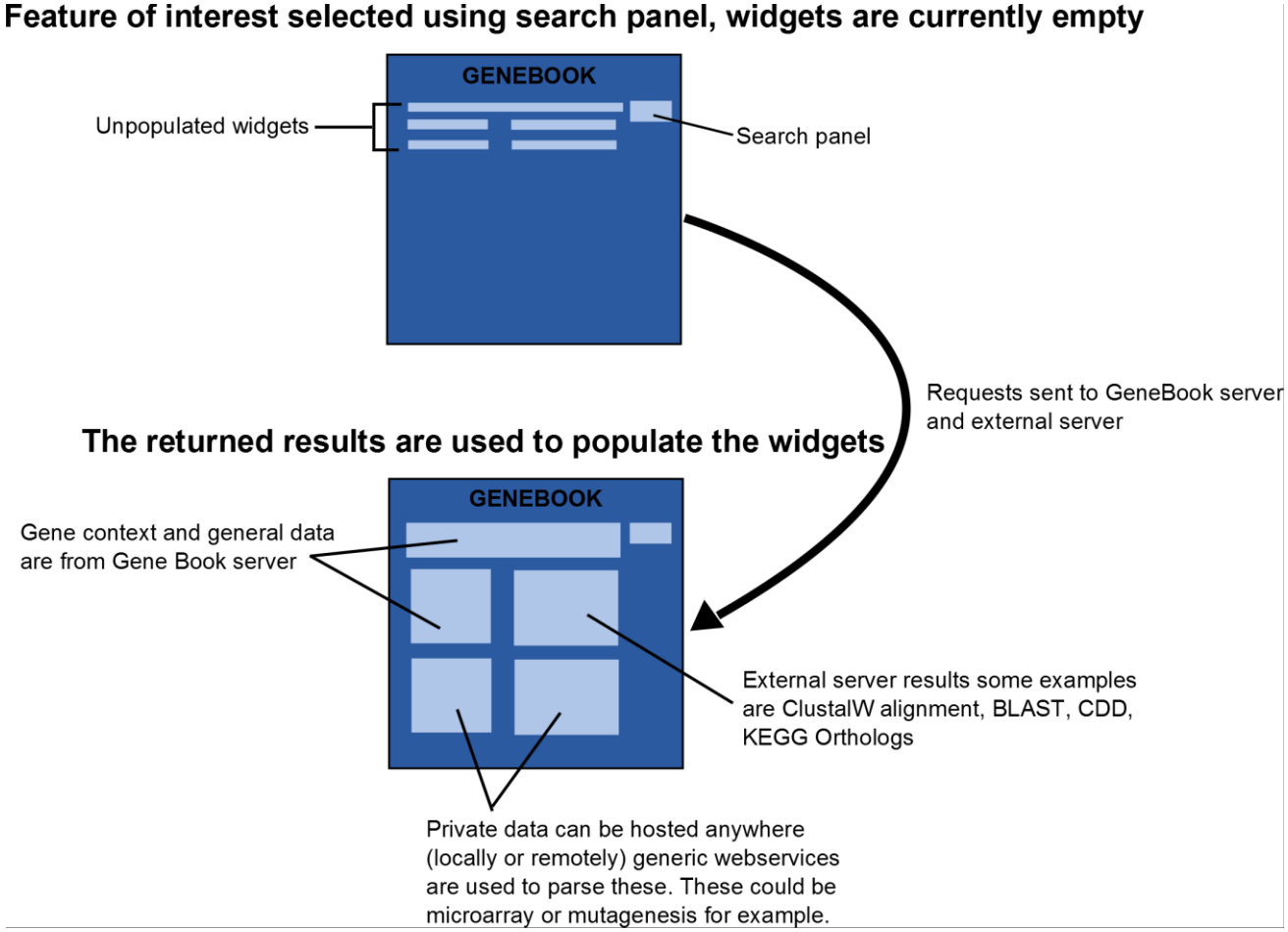


Figure 58 An overview of how GeneBook displays its data, showing before and after the user has selected a feature of interest.

4.3.1 Overcoming errors, outdated annotation

The proliferation of errors is an on-going phenomenon, as new genomes are sequenced there will be always be continuity issues as described in section 2.3.2.1. In most genome databases the data is static, over time this will become out of date, with experimental discovery and as our understanding of bacterial genetics improves.

This section demonstrates how GeneBook can overcome some of these inconsistencies by displaying up to date information in one browser window.

4.3.1.1 Hypothetical proteins

When a genome is first annotated, its open reading frames are aligned and compared to publicly available sequences using tools such as BLAST [79]. Many of these potential genes only hit other hypothetical proteins, which indicates sequence conservation but does not give any clues to functionality. After submission the proteins are automatically submitted to Uniprot. Uniprot is constantly being updated as new information about proteins is discovered. With that in mind it is possible that the original annotation is still labelled as a hypothetical protein but the corresponding annotation in other databases, such as Uniprot, show more informative, up to date annotation.

Using GeneBook to look at features of interest can save the user the effort of searching across multiple websites and having multiple tabs open. Integrating one's personal data with the public data can add evidence to a sequence feature actually being functional rather than being a false positive from feature prediction software such as GLIMMER [59].

For example, STM0081 is described as hypothetical protein. When it is viewed in GeneBook the BLAST widget shows the top UniprotKB hit is to:

Q8ZRW6_SALTY Putative secreted protein OS=Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)

This is a hit to itself, but with a more informative annotation. The TraDIS data further supports the inference that this is a functional protein (Figure 59). The figure

shows that STM0081 is negatively selected (with p-values less than 0.05) across all hosts meaning that this is needed for non-systemic infection.

When looking at STM0081 in the genome context widget we can see that its neighbour STM0082 is also a hypothetical protein Clicking on the feature opens a new tab and displays the integrated information for this protein. As with STM0081 there is a more up to date annotation in the BLAST results against UniprotKB:

Q7CR88_SALTY - Putative secreted protein OS=Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)

The Conserved Domain Database (CDD) widget also offers some more information on the protein sequence. Amongst the DUF and hypothetical proteins is a hit against:

PRK14864 putative biofilm stress and motility protein A; Provisional

This description could explain why STM0082 is significantly up regulated in the Macrophage (mouse macrophage) data (Figure 60) as biofilm formation has been linked to intracellular proliferation [233].

Both of these proteins are up to date in Xbase and PATRIC, in the respect that they have the most recent Uniprot annotation. However the addition of the most up-to-date BLAST results, CDD alignments and the private data sheds more light on these proteins' behaviour in defined conditions and perhaps helps users identify genes of particular interest to their studies.

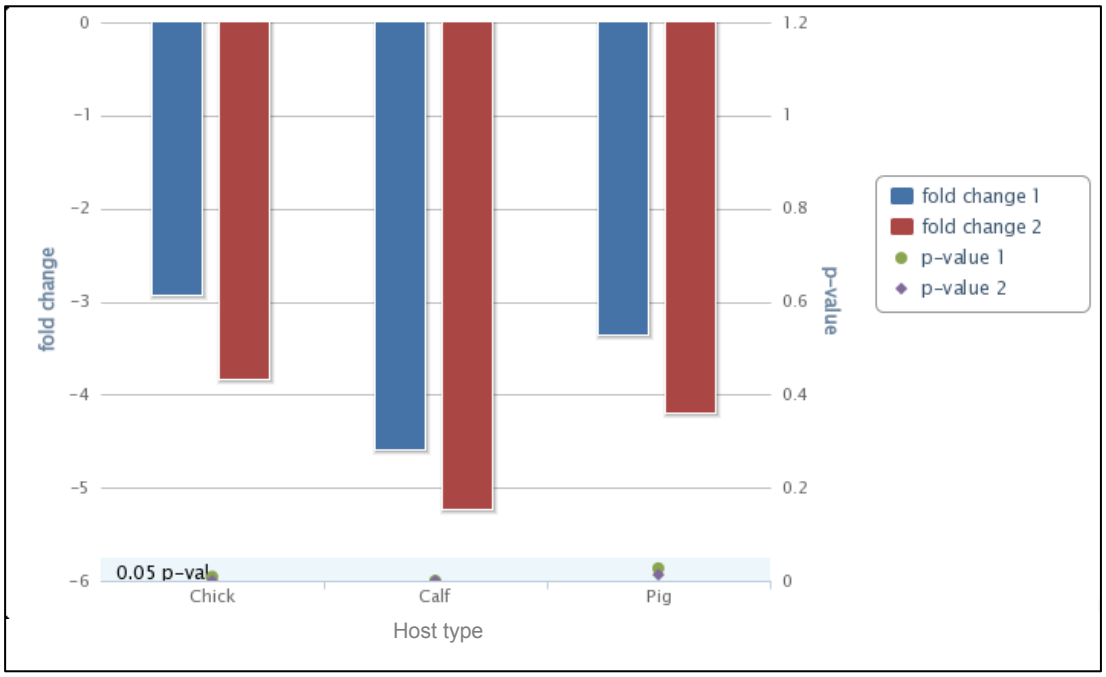


Figure 59 TraDIS data for STM0081 is significantly negatively selected across all hosts suggesting that it is essential for intestinal colonisation. The user can see that the results are all significant because the p-values fall within the light blue band.

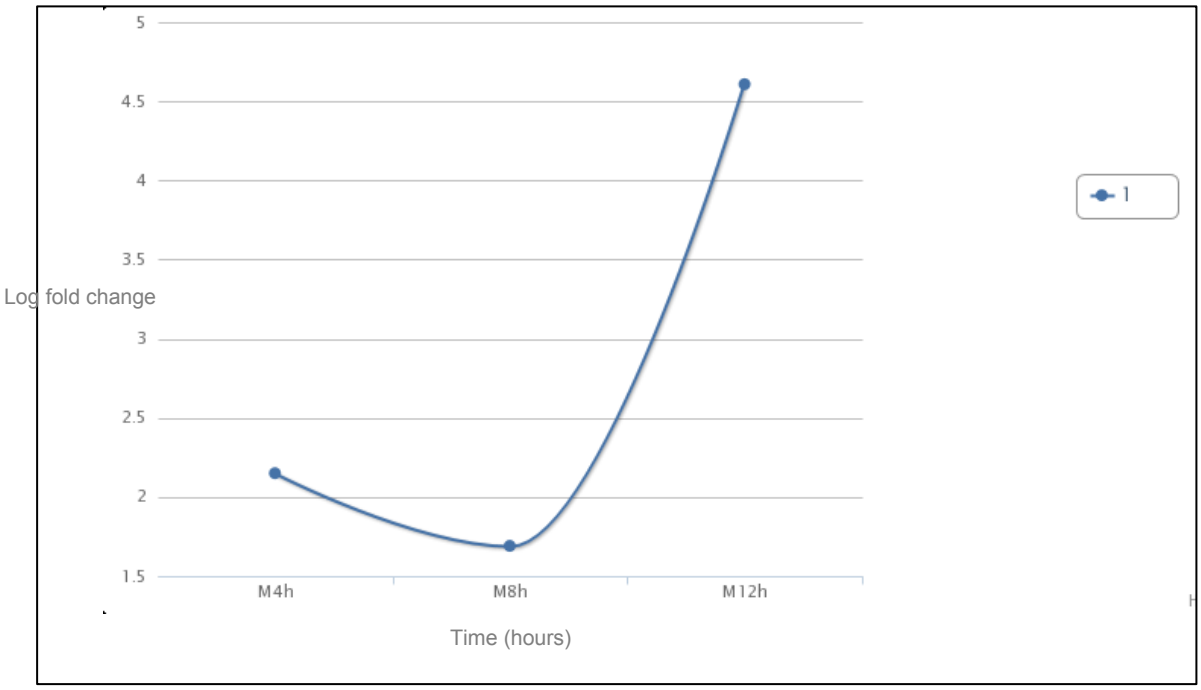


Figure 60 Macrophage data for the 'hypothetical protein' STM0082, it is up regulated suggesting activity during macrophage infection.

It is worth noting that Xbase and PATRIC are not always completely up to date. As an example of this, STM2220 is a 'hypothetical protein' in the original genome annotation and in both the previously mentioned genome browsers. However, in GeneBook the BLAST results return:

Q7CQ73_SALTY Putative cytoplasmic protein OS=Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)

There is also an interesting hit to the CDD using the eutils app:

pfam06952 PsiA This family consists of several Enterobacterial PsiA proteins. The function of PsiA is unknown although it is thought that it may affect the generation of an SOS signal in Escherichia coli

These two hits give some information above and beyond the genome annotation and the two main publically available bacterial genome browsers. Arguably, with their next synchronisation with Uniprot and other resources Xbase and PATRIC will be up to date again but until they are in sync with real time data their users will be unaware that this is out of date unless they visit other websites or use search tools to ensure they are receiving the most relevant information. BLAST is available on Xbase but as this is performed in a different window/tab on the browser it could prove confusing if the user has multiple tabs already open. GeneBook displays the BLAST results automatically in the same page as the rest of the data making it completely clear what results one is looking at in relation to a specific feature.

As an aside, another possible indication that STM2220 is a truly functional protein and not a relic from the annotation process is that according to the widget which holds the macrophage data (Figure 61) the protein is up regulated during macrophage infection (compared to control). This might be indicative of the gene's involvement in pathogenicity.

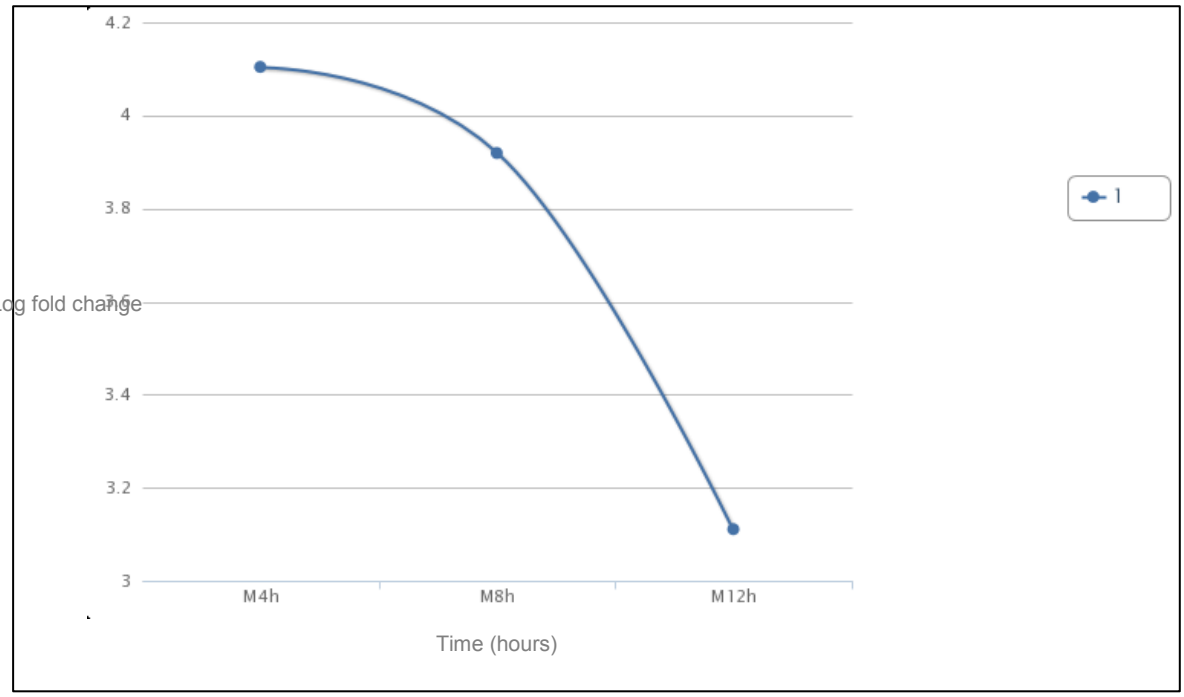


Figure 61 Macrophage data for STM2220, it is up regulated compared to the control but decreases over time.

4.3.1.2 Missing connections such as orthologs

Data for orthologous genes is available in resources such as COG and KEGG. KEGG has an API allowing straightforward access to the gene information, including orthologs and pathways. A recurring theme throughout this project is that using the information gained through data sources relies on the assumption that the data is correct and up to date. KEGG contained less than ten *Salmonella* genomes at the time of writing. Each genome within KEGG has had its genes mapped to pathways. With that in mind it is a fair assumption that if a pathway is detected in one gene then it should as easily be detected in any equivalent genes.

In GeneBook when the user selects STM1958 the KEGG ortholog browser widget returns one ortholog SC1961, Choleraesuis SC-B67 (Figure 62). The BLAST table widget shows that the other genomes do have this gene (Appendix A: STM1958_BLAST.xlsx).

Using the genome context widget to look at the neighbouring gene STM1960 one can see that the fliB is present across all the genomes (Figure 63), hovering over the respective genes shows that they are all labelled as fliB, although their KEGG entries do not have any orthology or pathways assigned to them (**http://www.genome.jp/dbget-bin/www_bget?see:SNSL254_A2119**).

On closer inspection, using the ClustalW widget we can see an alignment across this region. Figure 64A shows that the centre of the fliB gene is highly variable. It is possible that due to the variable region in the middle of the gene sequence the similarity threshold wasn't low enough to recognise these other orthologs. High variability can be linked to adaptation to the host immune system [234]. Interestingly this gene has been suggested to influence pathogenicity, in their 2006 paper Frye *et al.* discuss the possibility of n-methylation of the lysine on flagella being linked to host evasion as Typhimurium is still motile when the fliB gene is mutated [235]. To further support this, the TraDIS widget shows significant negative selection across all hosts (Figure 65). Conversely the Macrophage data shows that the gene is significantly down regulated over time (Figure 66), suggesting that when infecting

systemically Typhimurium uses different methods of host evasion or at least doesn't require flagella expression during the entire course of infection.

This example demonstrates that firstly relying on one resource for information is not failsafe, there can be errors. Secondly, relying on sequence homology alone is not always enough to find commonality between genomic features. By looking at the genome synteny, that is the pattern of the features around our region of interest, we can see that STM1960 has orthologs in other genomes and also a paralogous gene within its own genome. The method used in GeneBook to determine this isn't ideal but there might be scope for developing a genome context widget that does not rely on sequence homology. Turning homology based on genome context information into annotation is not a simple feat but would overcome these discrepancies.

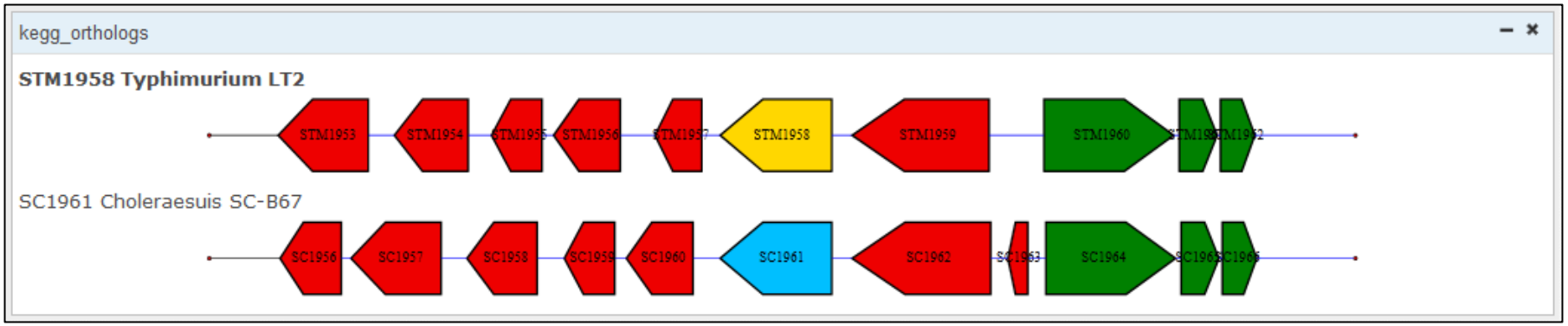


Figure 62 Genome context for STM1958. The KEGG API only returns one ortholog, SC1961 in Choleraesuis SC=B67

Figure 63 Genome context for STM1960. Using this gene to align the orthologous regions in other genomes it shows that there are more orthologs in STM1958 than the KEGG API returns (including three pseudogenes).

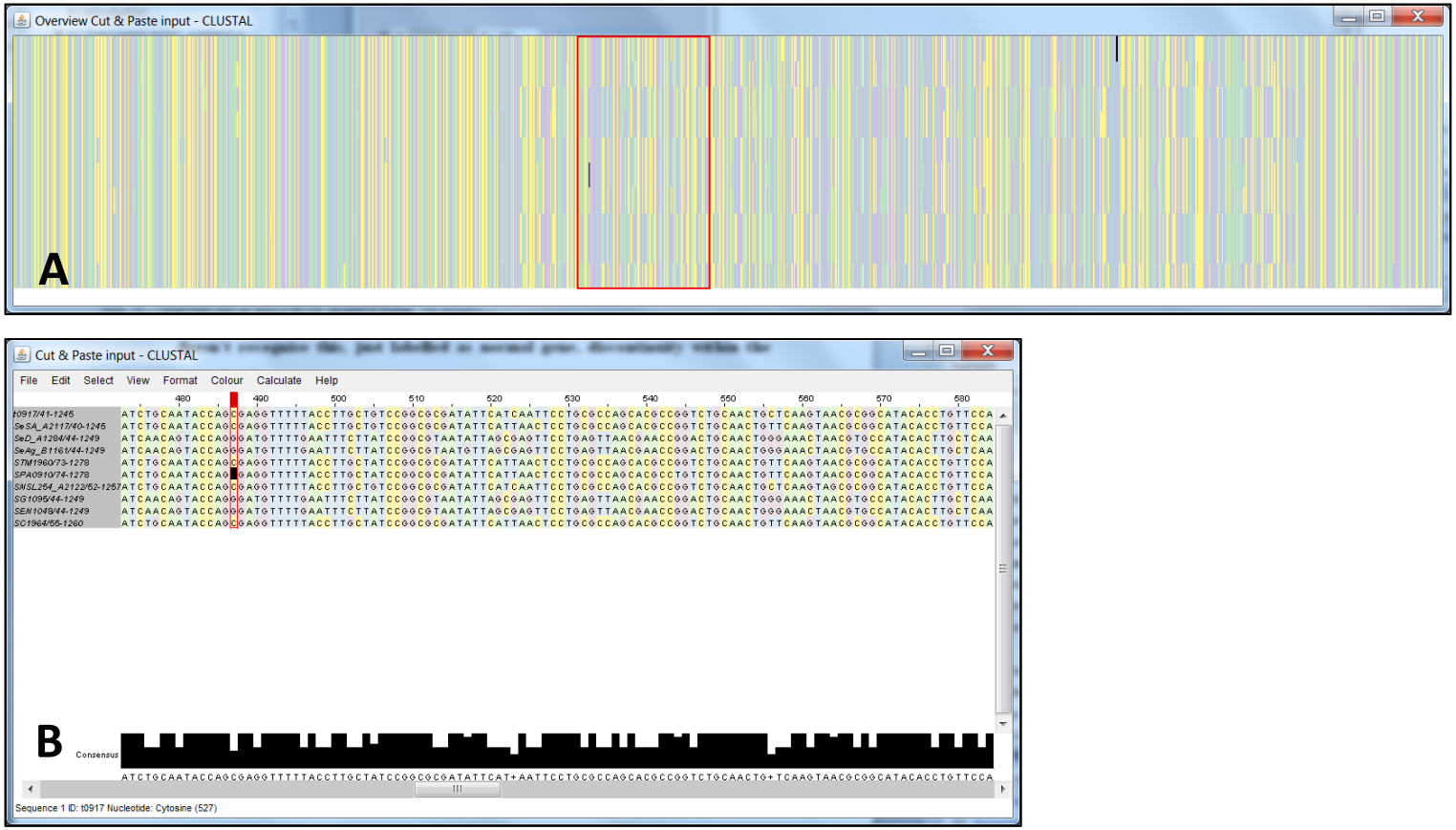


Figure 64 The alignment of STM1958, taken as a subsequence of the STM1960 (with 3500bp upstream and downstream (A) An overview of the alignment, black areas show gaps in the alignment. Vertical lines of the same colour show conservation, disjointed colours (such as the area in red) indicate sequence variation between genomes. (B) Nucleotide level view of the alignment in a highly variable region. The region of nucleotide deletion in SPA0910 is highlighted in red with the actual deletion shown in black.

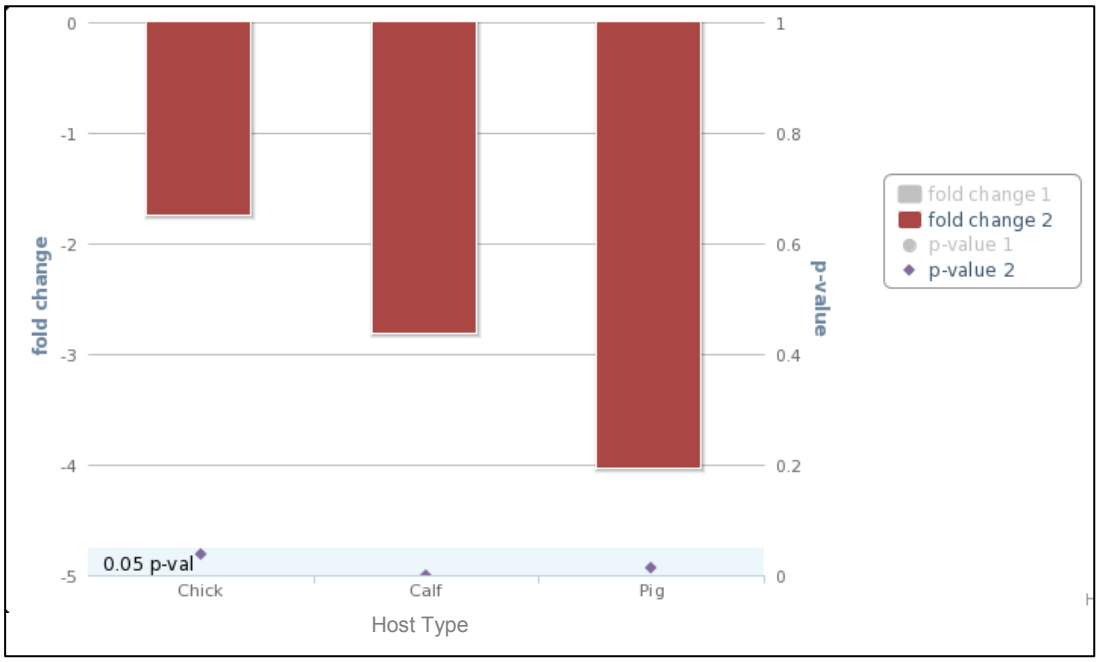


Figure 65 TraDIS data for STM1958 showing significant negative selection across all hosts.

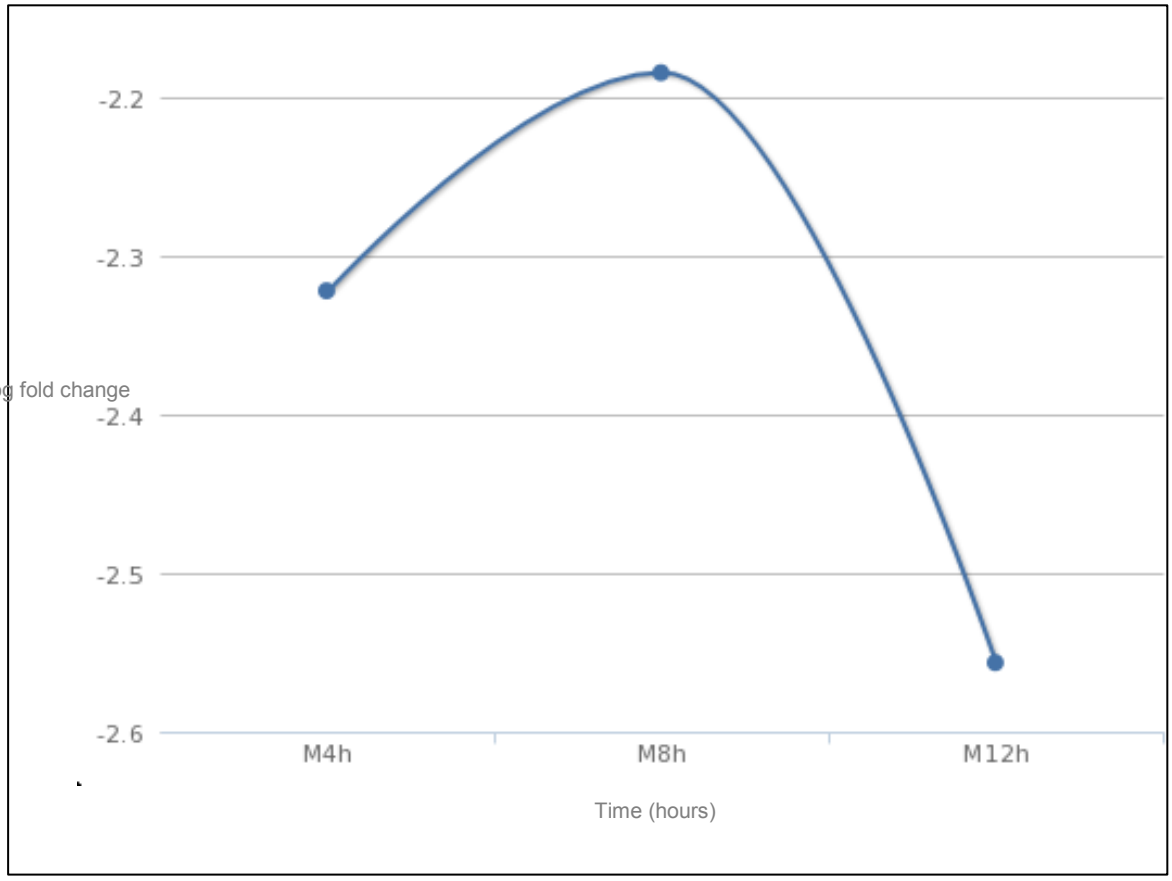


Figure 66 Macrophage data for STM1958 showing significant down regulation of the gene at each data point.

4.3.1.3 Uninformative and common gene names

Gene names change as we find out more about their behaviour. For example, y-genes which were discussed in section 2.4.2.3 are hypothetical proteins, whose gene name is specific to the location on circularised *Escherichia* coli *K*-12 [158]. These names passed across to other serovars and even other species. For example STM0246 is tagged as having the gene name *yaeE* this has no context in *Salmonella* Typhimurium LT2 and gives no information regarding function. The protein description in xBase, PATRIC and GeneDB is:

yaeE - DL-methionine transporter permease subunit

GeneDB has noted an ortholog in the comments:

Ortholog of E. coli yaeE (YAEE_ECOLI); FASTA hit to YAEE_ECOLI (217 aa), 93% identity in 217 aa overlap

However, those E.coli orthologs no longer exist, when searching for them in Uniprot one is taken to the more up to date entry (below). The KEGG pathway widget displays a pathway for this gene with the other LT2 genes in that pathway listed below the pathway description (Figure 67). When the user clicks on the link the image is dynamically displayed over the browser (Figure 68). The selected gene is automatically highlighted in red so that the user can clearly see where it lies in the pathway. In the case of STM0246 the gene in the pathway is described as MetI (Figure 69), showing a more up to date gene name. Further to this, the BLAST widgets support the new gene (Figure 70 and Figure 71). The BLAST table widget (Table 26) shows hits to TREMBL and Swissprot with the top hit being:

METI_SALTY D-methionine transport system permease protein metI

Finally, the table version of the KEGG orthology webservices shows that there is an in-paralog within LT2 and that there are orthologs and out-paralogs in the other genomes Table 27. It is also worth noting that the gene name and product names for the homologues are varied, emphasizing the varying quality of bacterial genome annotation.

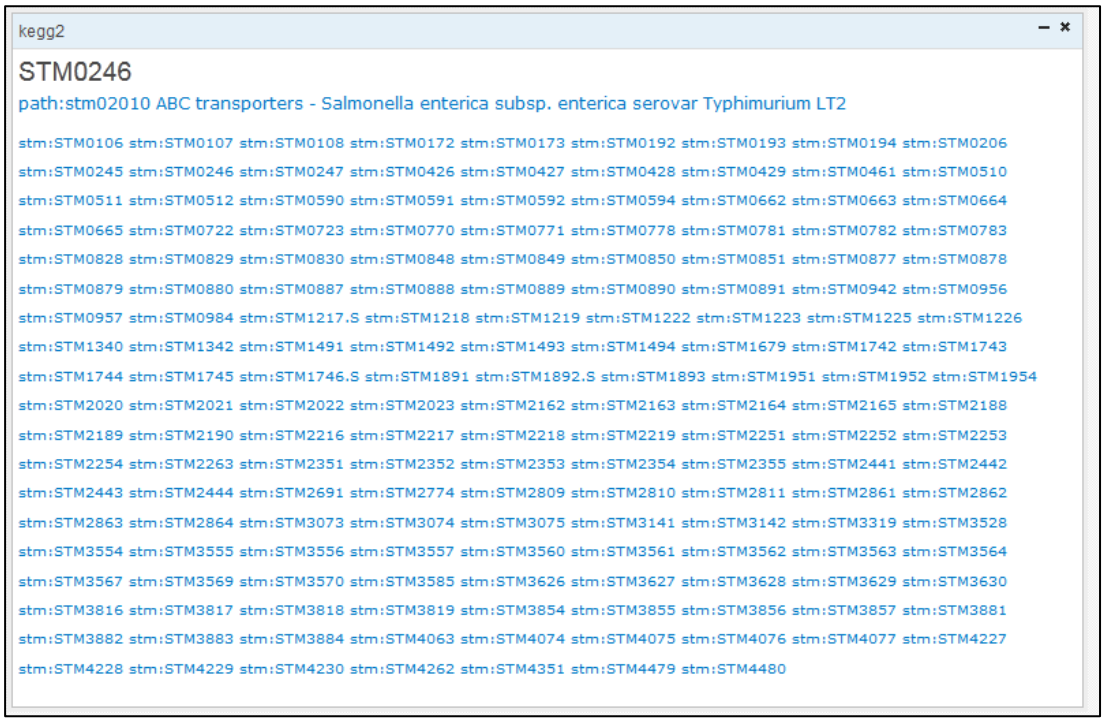


Figure 67 KEGG widget showing the pathway hit for STM0246 and the other genes from LT2 that belong to this pathway. The pathway link opens the pathway map on top of GeneBook (Figure 68). The locus tag hyperlinks take the user to the respective entry in GeneBook.

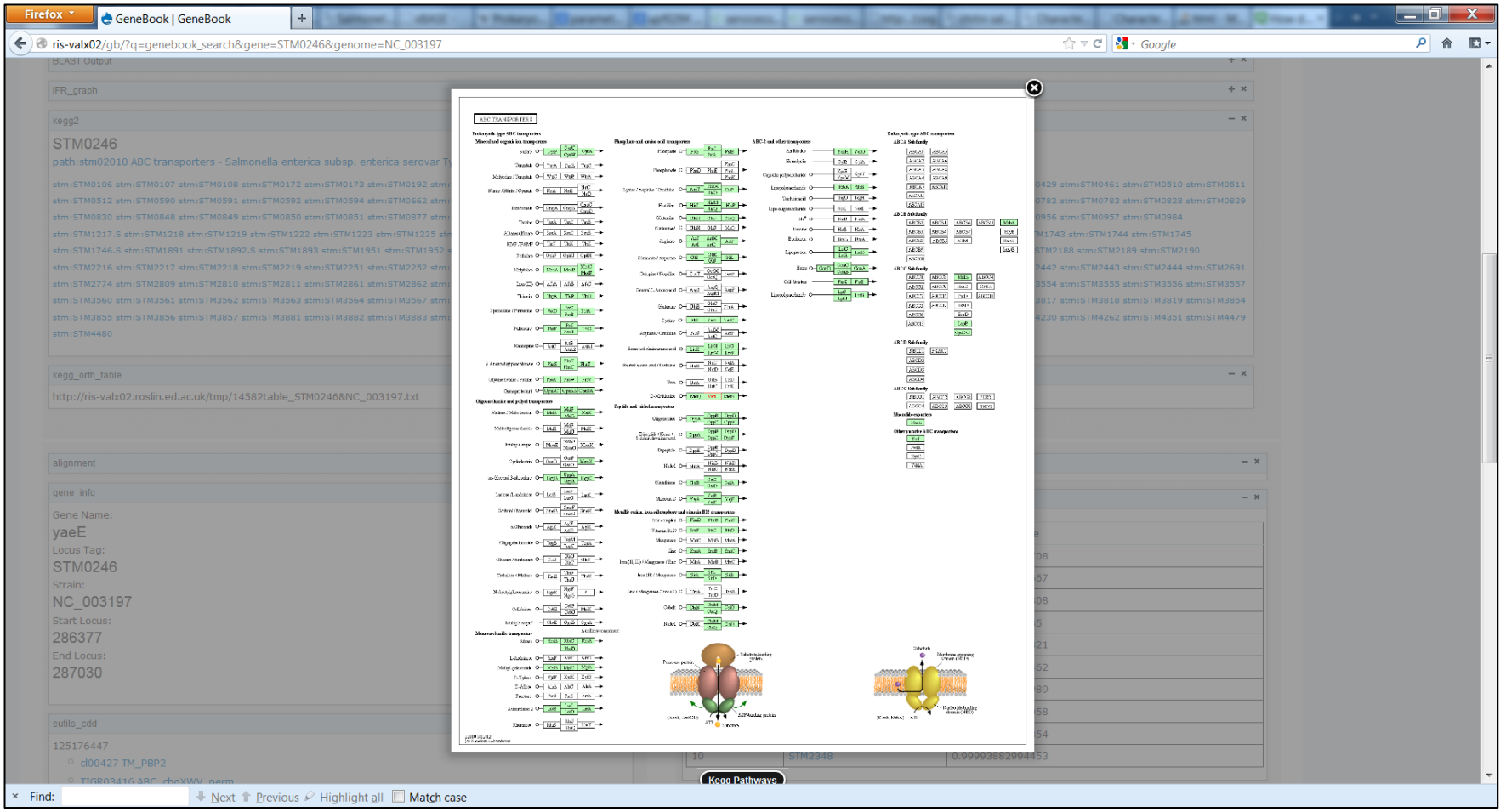


Figure 68 The pathway map for ABC Transporters displayed dynamically on top of GeneBook.

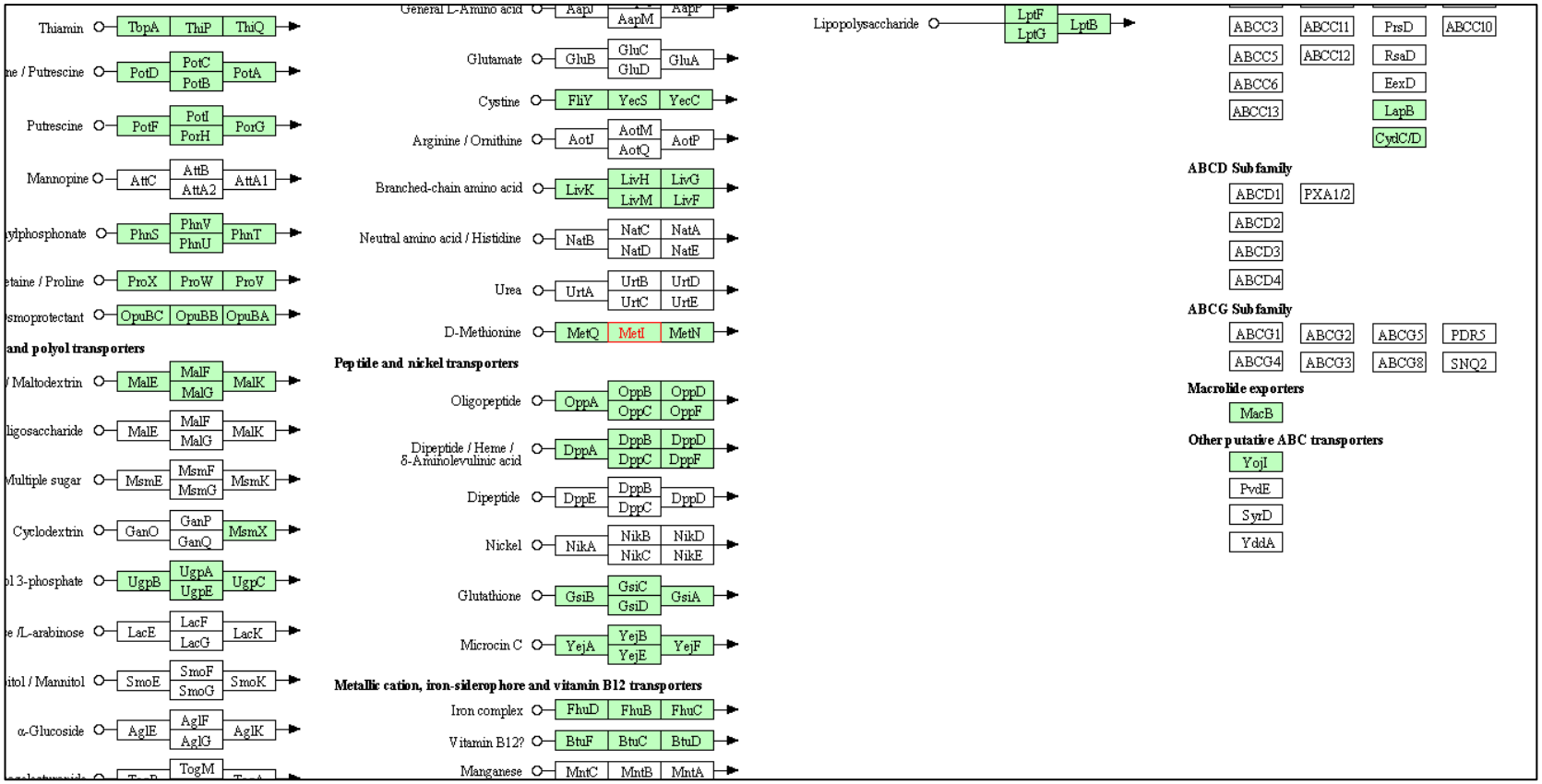


Figure 69 GeneBook output for STM0246 from the KEGG Map widget zoomed in on part of the STM0210 ABC Transporters pathway diagram. The location of STM0246 is highlighted in red showing that this is the metI gene rather than yaeE. GeneBook provides the link with the highlighted link for future use. (http://soap.genome.jp/tmp/mark_patway_www_api.13521279182058/stm02010.png)

Table 26 Top 5 BLAST hits to UniprotKB for STM0246 from the BLAST table widget. In the widget the Hit id is a hyperlink taking the user to the Uniprot entry

Hit id	Hit accession	Description	E-value	Length
METI_SALTY	Q8ZRN0	D-methionine transport system permease protein metI OS=Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720) GN=metI PE=3 SV=1	1E-148	217
Q57T10_SALCH	Q57T10	Putative ABC superfamily (Membrane) transport protein OS=Salmonella Choleraesuis (strain SC-B67) GN=yaeE PE=3 SV=1	1E-148	217
F5ZM31_SALTU	F5ZM31	DL-methionine transporter permease subunit OS=Salmonella typhimurium (strain ATCC 68169 / UK-1) GN=yaeE PE=3 SV=1	1E-148	217
E8XIR6_SALT4	E8XIR6	DL-methionine transporter permease subunit OS=Salmonella typhimurium (strain 4/74) GN=yaeE PE=3 SV=1	1E-148	217
E1W8B8_SALTS	E1W8B8	Hypothetical ABC transporter permease protein OS=Salmonella typhimurium (strain SL1344) GN=yaeE PE=3 SV=1	1E-148	217

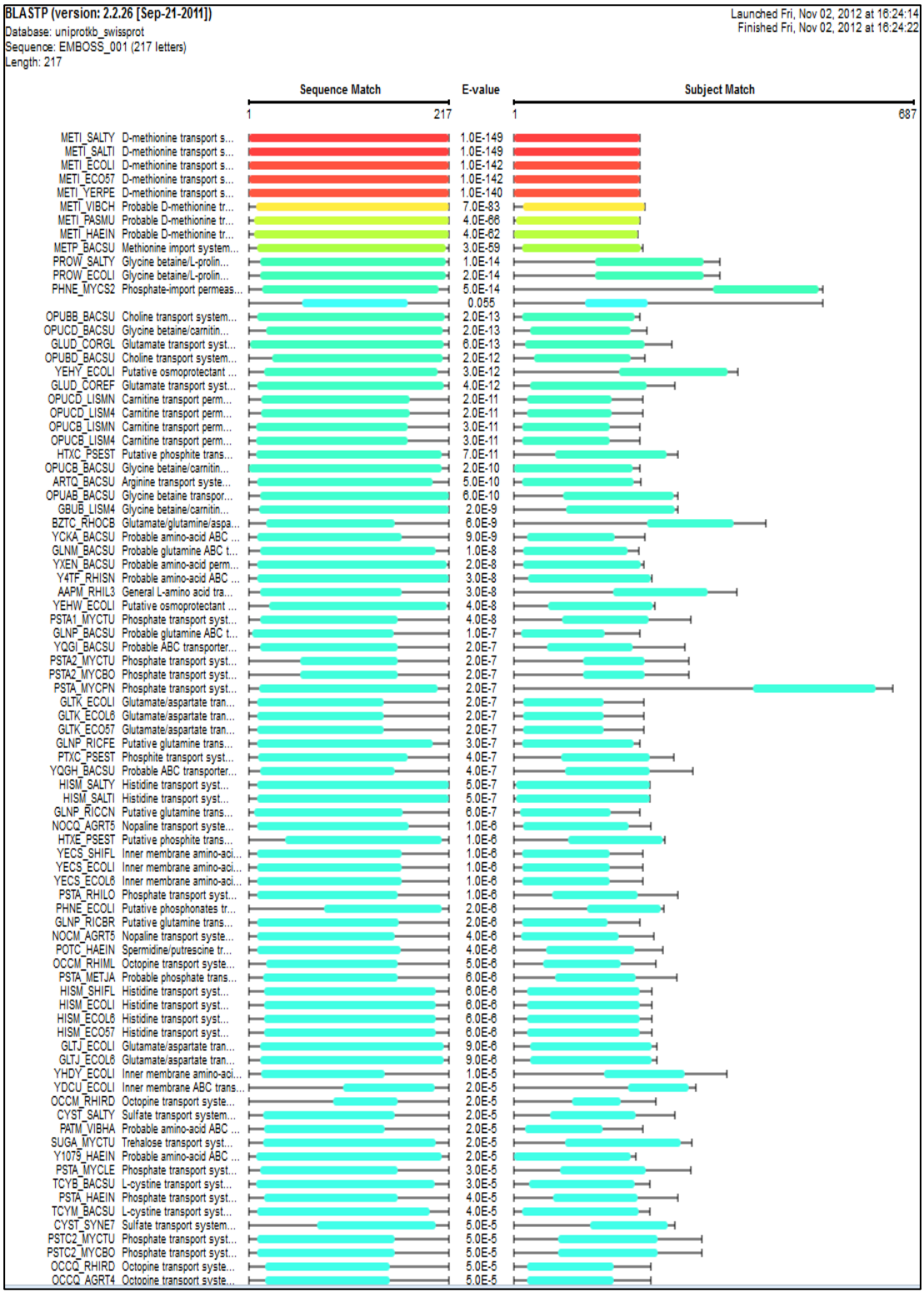


Figure 70 BLAST hits for STM0246 against Swissprot returns D-methionine transporter permease subunit as the top hit. The GeneBook output has hyperlinks embedded meaning that the user can click on the link and the Uniprot entry is returned in the widget (Figure 71)

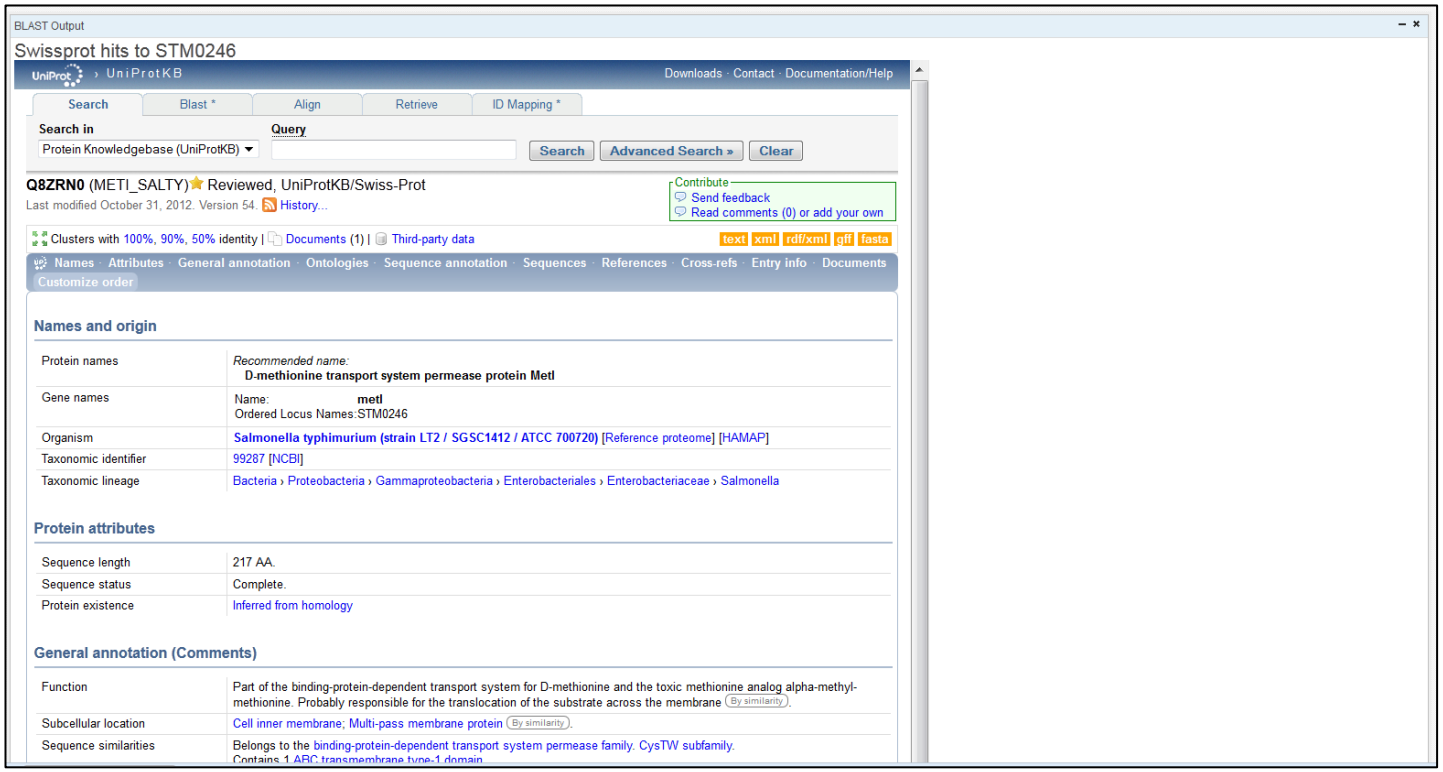


Figure 71 Shows the Uniprot data within the BLAST widget, this is loaded without affecting the other widgets and can be returned to the BLAST results by using the 'backwards' button in the web browser, which also does not affect the results.

Table 27 Output from the Table version of the KEGG ortholog widget. The putative in-paralog has been highlighted in red and the out-paralogs are italicised.

Locus tag	Gene name	Product	Serovar	Strain
STM0246	yaeE	DL-methionine transporter permease subunit	Typhimurium	LT2
SC0245	yaeE	DL-methionine transporter permease subunit	Choleraesuis	SC-B67
SPA0253	yaeE	DL-methionine transporter permease subunit	Paratyphi A	ATCC 9150
STY0273	yaeE	DL-methionine transporter permease subunit	Typhi	CT18
t0249	yaeE	DL-methionine transporter permease subunit	Typhi	Ty2
SeAg_B0287	metI	D-methionine ABC transporter, permease protein	Agona	SL483
SeD_A0268	metI	D-methionine ABC transporter, permease protein	Dublin	CT_02021853
SeHA_C0284	metI	D-methionine ABC transporter, permease protein	Heidelberg	SL476
SeSA_A0273	metI	D-methionine ABC transporter, permease protein	Schwarzengrund	CVM19633
SNSL254_A0268	metI	D-methionine ABC transporter, permease protein	Newport	SL254
SEN0254	yaeE	putative ABC transporter permease protein	Enteritidis	P125109
SG0250	yaeE	putative ABC transporter permease protein	Gallinarum	287/91
STM0512	sfbC	putative ABC transporter permease component	Typhimurium	LT2
SC0553	*sfbC*	*putative binding-protein-dependent transport systems inner membrane component*	*Choleraesuis*	*SC-B67*
SeAg_B0559		*ABC transporter integral membrane protein*	*Agona*	*SL483*
SeD_A0561		*ABC transporter integral membrane protein*	*Dublin*	*CT_02021853*
SeHA_C0620		*ABC transporter integral membrane protein*	*Heidelberg*	*SL476*
SEN0493	*sfbC*	*ABC transporter integral membrane protein*	*Enteritidis*	*P125109*
SeSA_A0576		*ABC transporter integral membrane protein*	*Schwarzengrund*	*CVM19633*
SG0523	*sfbC*	*ABC transporter integral membrane protein*	*Gallinarum*	*287/91*
SNSL254_A0566		*ABC transporter integral membrane protein*	*Newport*	*SL254*
SPA2210	*sfbC*	*ABC transporter integral membrane protein*	*Paratyphi A*	*ATCC 9150*
STY0560	*sfbC*	*ABC transporter integral membrane protein*	*Typhi*	*CT18*
t2348	*sfbC*	*ABC transporter integral membrane protein*	*Typhi*	*Ty2*

4.3.1.4 Annotation inconsistencies between genomes

The transferral of annotation from one genome to another relies on the assumption that the original annotation is accurate. The eutM/eutN example is explained in section 1.4.2. This model can be demonstrated in GeneBook. Selecting STM2464 shows that this is the eutN gene. This eutN description concurs with the original annotation, and the general consensus of the BLAST results and CDD show this too. However, if the user wanted to look at the ortholog SPA0404 (genome NC_006511 Paratyphi C) the annotation describes the gene name as eutM, the BLAST widget and BLAST ortholog widget (Table 28) both show that the closest relatives are eutN. This suggests that the gene name for SPA0404 is an error, presumably a typo.

The eutN/M example is interesting beyond the annotation inconsistencies, Figure 72 shows the variation between genomes. Two areas of interest are the eutN genes for the Typhi strains being considerably larger than the other eutNs and the eutMs in serovars Newport and Agona being labelled as pseudogenes.

Table 28 Orthologous hits to the eutN gene in Typhimurium, STM2464, from the KEGG API. The product descriptions are varied and the gene name for SPA0404 (highlighted and underlined) is eutM rather than eutN indication a possible annotation error.

Locus tag	Gene name	Product	Serovar	Strain
STM2464	eutN	putative detox protein	Typhimurium	LT2
STY2701	eutN	putative ethanolamine utilization protein eutN	Typhi	CT18
t0394	eutN	putative ethanolamine utilization protein eutN	Typhi	Ty2
SPA0404	**eutM**	putative ethanolamine utilization protein eutN	Paratyphi A	ATCC 9150
SC2460	eutN	putative detox protein in ethanolamine utilization	Choleraesuis	SC-B67
SNSL254_A2657	eutN	ethanolamine utilization protein	Newport	SL254
SeHA_C2724	eutN	ethanolamine utilization protein	Heidelberg	SL476
SeAg_B2609	eutN	ethanolamine utilization protein	Agona	SL483
SG2495	eutN	ethanolamine utilization protein	Gallinarum	287/91
SEN2444	eutN	ethanolamine utilization protein	Enteritidis	P125109

Figure 72 Genome context of eutN, the two Typhi genes are markedly bigger than the other orthologous genes.

4.3.1.5 Pseudogene assessment

Pseudogene detection, definition and determination could be a thesis in itself. Trying to identify a true pseudogene is no mean feat. For example, a gene that has an indel which has led to a frameshift could still be functional. Transcription of a given gene doesn't necessarily confer function either, some pseudogenes are still transcribed, they just don't make a functional protein. By definition a pseudogene is a gene that is no longer functional due to random mutations. These mutations can result in: a frameshift, disrupting the gene; an introduced stop codon, truncating the gene; a different three dimensional structure that renders the protein useless. The final example couldn't be conclusively proved as a pseudogene but exploration for non-synonymous SNPs would identify the mutation could form a basis for functional studies. *In silico* pseudogene is possible for the first two examples. Scrompt-Ruttedge do this by taking hypothetical proteins (on the basis that they might be truncated genes) and intergenic regions (possible parts of pseudogenes that are no longer in an open reading frame). They compare these regions to Uniprot, when they get a hit they look for stop codons/indels in the flanking regions [Liu 2004].

Trusting the annotation from a reference genome can lead to the misannotation of pseudogenes. The eutM/N example can be used to demonstrate how the user can explore pseudogenes and assess the annotation for themselves. Figure 72 shows the discrepancies between different genomes around eutN/M highlighting a couple of questions:

- Is the eutN/M model incorrect, should it just be eutN and those which show two open reading frames are pseudogenes?
- Is the Typhi eutN gene a merged eutN and eutM as a result of a frameshift resulting in one open reading frame?
- Are the eutMs in Newport and Agona labelled as pseudogenes because their annotation was based on the eutM/eutN model and they have a mutation in eutM?

GeneBook was used to look at eutN in Typhi (t0394). The 'alignment with buffers' widget shows the ClustalW alignment with 1500 bases added either side of the orthologs. The extra bases allow the shorter orthologs and eutM to align to the larger Typhi eutN which spans both. Figure 73 shows that there is a deletion in t0394 relative to its orthologs which has resulted in a frameshift that merges eutN and eutM into one open reading frame.

eutM, Newport SNSL254_A2658 and Agona SeAg_B2609 are both annotated as pseudogenes. Using the 'alignment with buffers' widget and highlighting areas of conservation we can see how these eutMs differ from the consensus (Figure 77). There are no indels relative to the other eutM genes in other genomes but there are two substitutions that are shared only in Agona, Newport and Typhi (Figure 74). A substitution alone wouldn't be grounds for annotating the genes in Agona and Newport as pseudogenes. According to the neighbour joining tree made by the widget the large Typhi open reading frame is most similar to Agona and Newport which are identical to one another (Figure 75). The GenBank annotation for these two 'pseudogenes' is the same too:

"ethanolamine utilization protein eutM; this gene contains a frame shift which may be the result of sequencing error; identified by match to protein family HMM PF00936"

This annotation does not concur with the eutN/M model as there is no apparent frameshift. If, however the annotation is based on the Typhi eutN/-M model then there is a frameshift relative to eutN in Typhi. If this is the case and Agona and Newport are pseudogenes due to an insertion relative to t0394 then all of the other ortholog should be labelled as pseudogenes too. Unfortunately, as the GenBank annotation of these pseudogenes does not state what protein they are similar to it is very difficult to determine which is the correct model.

Figure 73 The ClustalW alignment of t0394 eutN. The deletion in t0394 which leads to a frameshift is present at the 278th base. This deletion is just within the eutM according the ortholog in Typhimurium (highlighted in red).

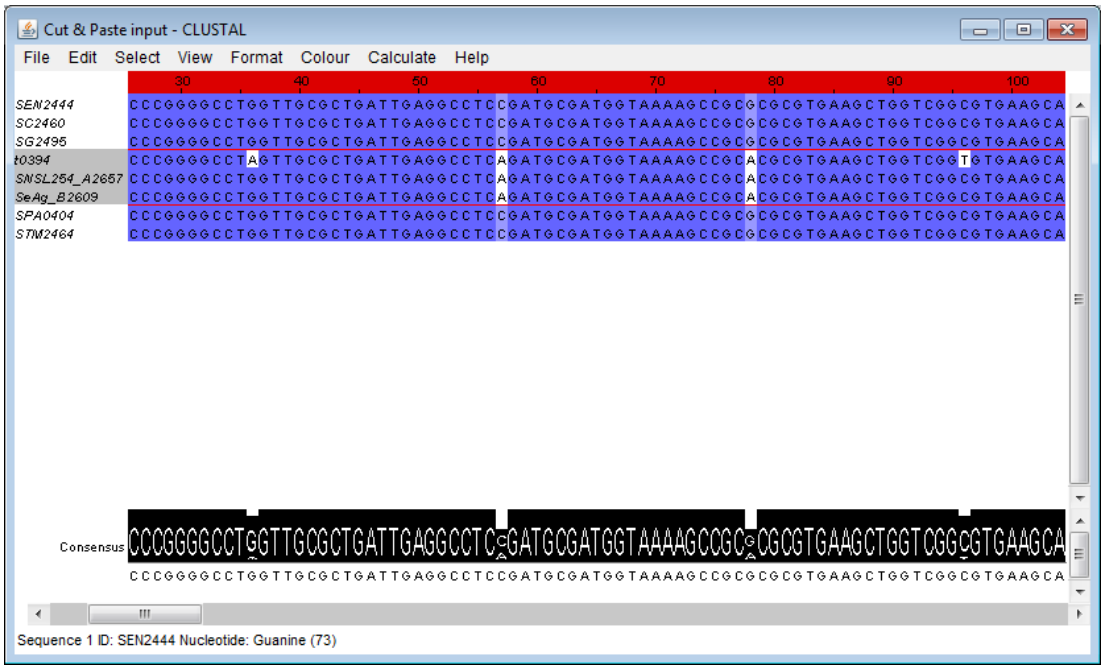


Figure 74 ClustalW alignment from alignment with buffers widget. The alignment is coloured by conservation, it highlights the two substitutions that make the Agona and Newport eutM different from the consensus sequence.

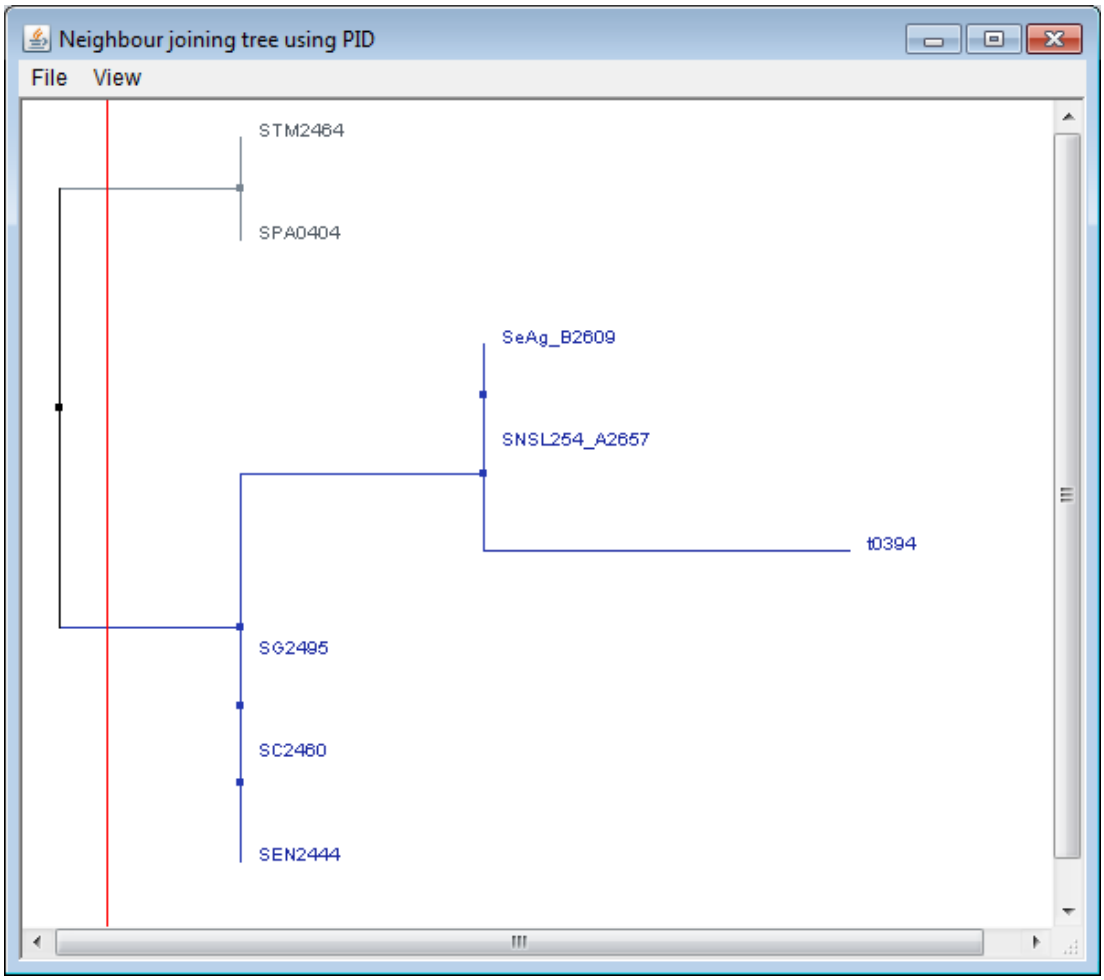


Figure 75 Neighbour joining tree of the sequence spanning Typhi eutN made by the alignment with buffers widget, showing that there is a lot of sequence similarity between genes, with the pseudogenes SeAg_b2609 and SNSL254_A2657 as identical and most closely related to the large eutN gene.

Another example of pseudogene interpretation is the Copper homeostasis protein cutF gene (Figure 76). This is a pseudogene in Choleraesuis SC-B67 (SCPS46) and Typhimurium LT2 (STM0241). Looking at the pseudogene in GeneBook will not return an alignment because the alignment is based on orthology from KEGG, which does not include pseudogenes. The alignment with buffers widget allows the user to look at the gene next to the pseudogene of interest as the extra bases (padding defined by the user) can span the pseudogene. The user can then isolate the region which specifically covers the pseudogene and view it separately.

The alignment shows that there is a deletion of 'ATGT' in Choleraesuis and an insertion of 'ATGT' next to this in Typhimurium (Figure 77). Closer inspection shows that in fact all of the orthologs across all of the *Salmonella* genes have this 4-mer, it is present once in this region in Choleraesuis, three times in Typhimurium and twice in the rest. With that in mind it is a big coincidence for these 4-mer indels to occur in a repeat region, this suggests that there may be a sequencing error rather than any pseudogenes.

GeneBook can help users to decipher the annotation around pseudogenes. Currently there is no silver bullet to predict pseudogenes, there are no hard and fast rules. Looking at the annotation in the context of multiple pieces of information empowers the user to decide the accuracy of the annotation rather than taking the annotations in public databases at face value.

In summary Section 4.3.1 shows how GeneBook is able to augment annotations that have little information by combining it with remote data sources as seen in section 4.3.1.1 and 4.3.1.3. Further to this section 4.3.1.2 demonstrates that GeneBook can identify errors in other data sources. Integrating multiple data sources gives users a fuller picture and allows them to make judgements based on all the data provided.

Further to this GeneBook is able to solve some of the problems of the eutN/M locus described in 1.4.2. The user can see their gene in context of all other genomes, and they can see live BLAST results alongside the ClustalW alignments.

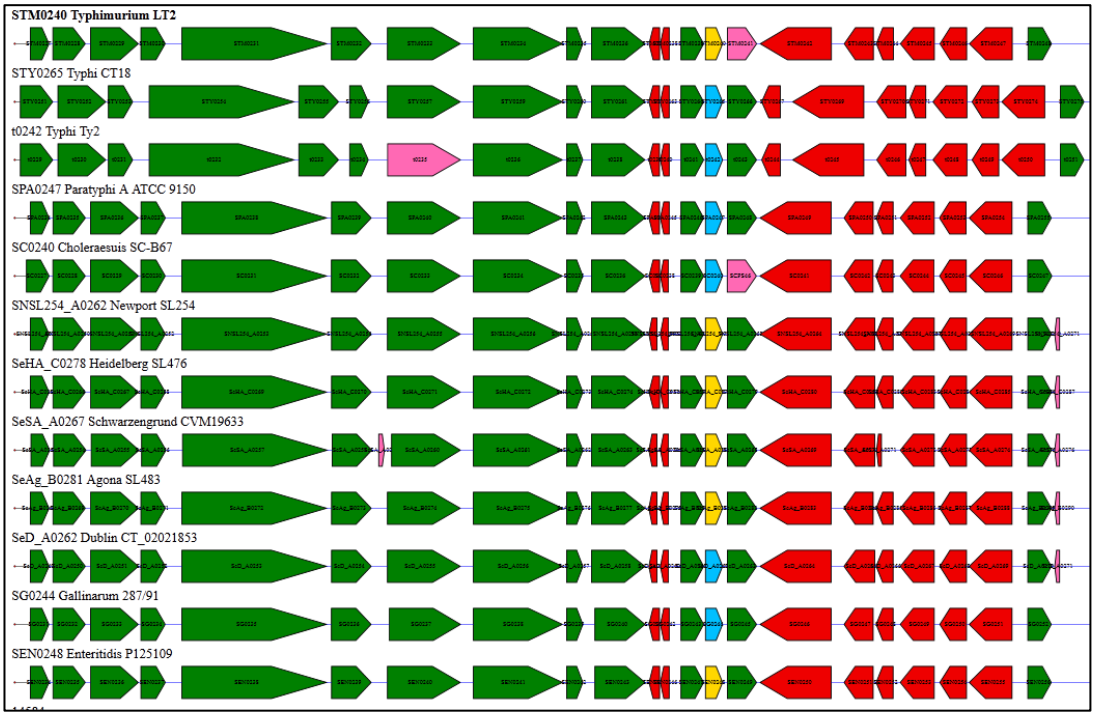


Figure 76 genome context diagram of STM0240 with orthologous genes from other serovars, based on KEGG orthology. This serves as a method for seeing the neighbouring pseudogene STM0241 alongside its orthologs, as KEGG does not calculate orthologs for pseudogenes.

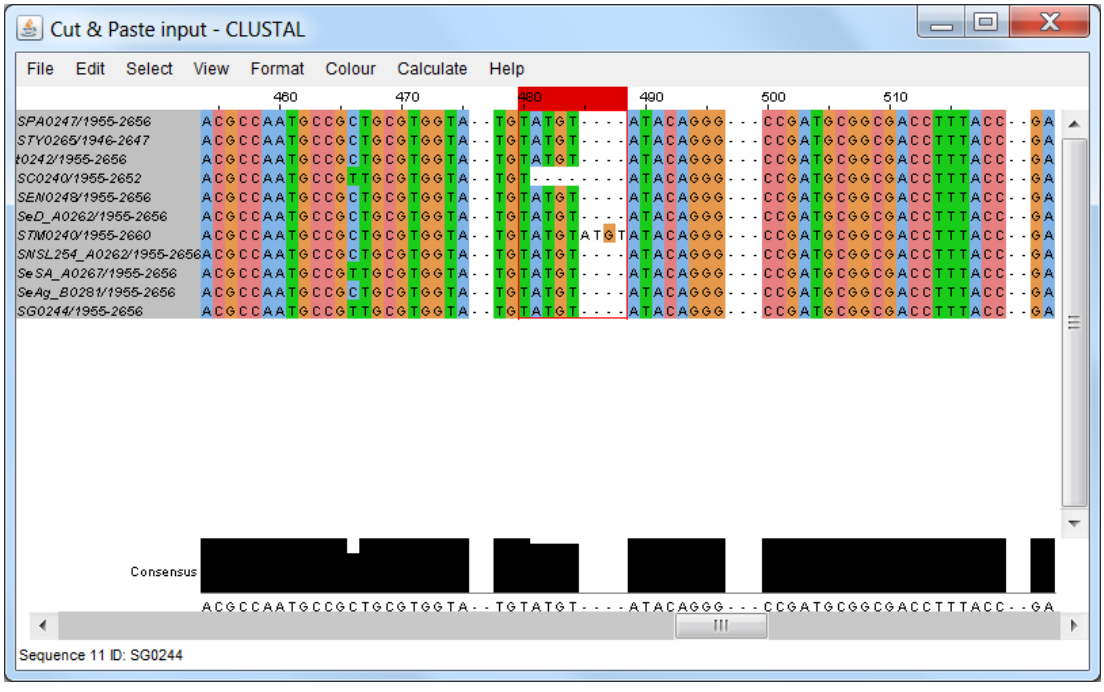


Figure 77 Subsequence from the ClustalW alignment of STM0240 with 1500bp up and downstream, specifically the pseudogene sequence of STM0241. The insertion in Typhimurium LT2 and deletion in Choleraesuis SCA50 are highlighted in red.

4.3.2 *Salmonella* growth in different conditions

Often, scientists have multiple data sets from the same organism. It would be useful for them to view the data in context of each other. It is possible to parse/merge these data but this becomes complex when there are more than a couple. Allowing users to integrate their own data into GeneBook, means that they can try to piece together a story. For example, the user may have a list of significant genes from a microarray, they can simply click on these genes in GeneBook and see how they are acting in other experiments and what data is publically available too.

This section shows some examples of how GeneBook can be used to augment the scientific meaning to genes of interest taken from the data explained at the beginning of this chapter. In these examples the 'user' has a list of gene of interest which are viewed in GeneBook to get a clearer picture of what they are and how they behave. In this scenario the user has a list of genes which they discovered to be significantly up regulated in the Macrophage data. That is the genes are up regulated in chicken macrophage when compared to broth.

When looking at the significant gene list in GeneBook, some genes showed up regulation in the Macrophage data too. These genes are being highly expressed in both extracellular (Caecum) and intracellular (Macrophage) colonisation relative to the control.

For example, STM0018 shows up regulation in both Caecal and Macrophage data (Figure 78 and Figure 79). The user can see both these widgets in the context of one another. Looking at the TraDIS data shows that the one of the point mutations is significantly negatively selected in calves and pigs (Figure 80). As there are four different point mutations and only one shows any significance, the user might want to see the point of this mutation. Looking at the context widget reveals the location of the mutation of interest (Figure 81). A simple domain search reveals that mutation is not in any domains.

The BLAST results return this as a ‘putative exochitinase’ and the eutils widget which describes the associated domains and functions (Appendix G), concurs with this stating that the protein is an exochitinase. The question that arises from these findings is; what is the relevance of chitinase in extracellular and intracellular infection? There is evidence in the literature that ChiA is up regulated in macrophage infection (Eriksson *et al.* 2003 in [236]). Further to this in 2010 Larsen *et al.* try to verify the role of exochitinase in *Salmonella* as chitinases are uncommon in bacteria [236]. Their findings showed enzyme activity towards both chitin and LacNac, which they state can be related to pathogenic strains that bind to the human intestinal epithelium (Humphries *et al.* 2009 in [236]).

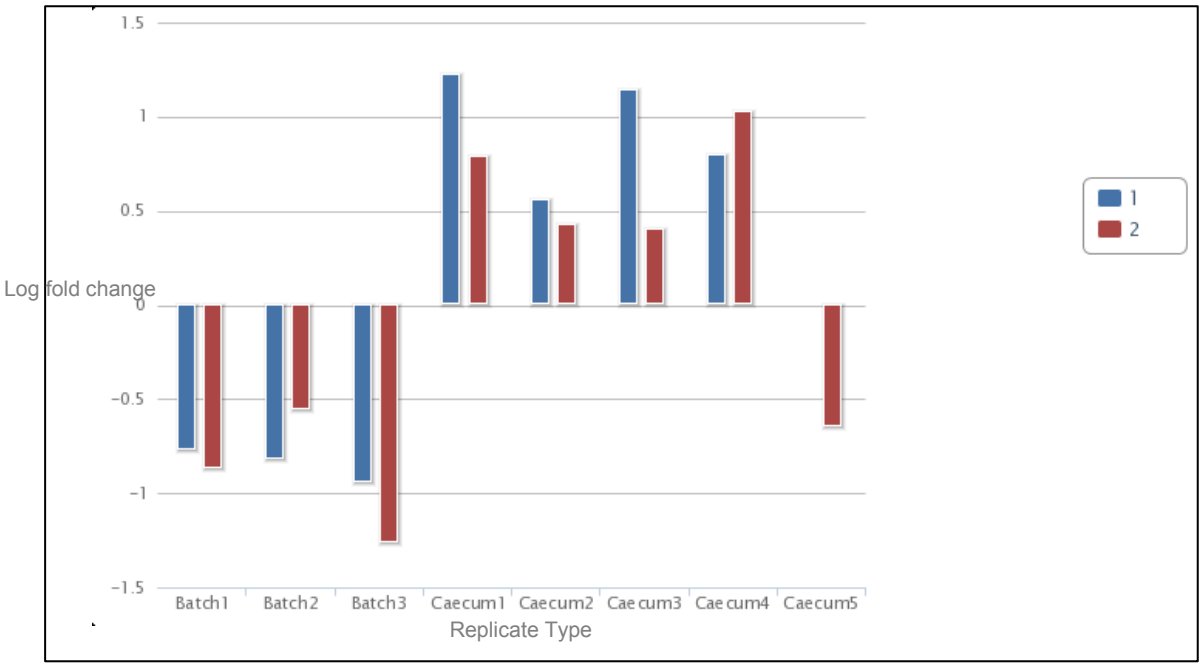


Figure 78 Caecal microarray data for STM0018, showing significant up regulation in caeca compared to LB Broth.

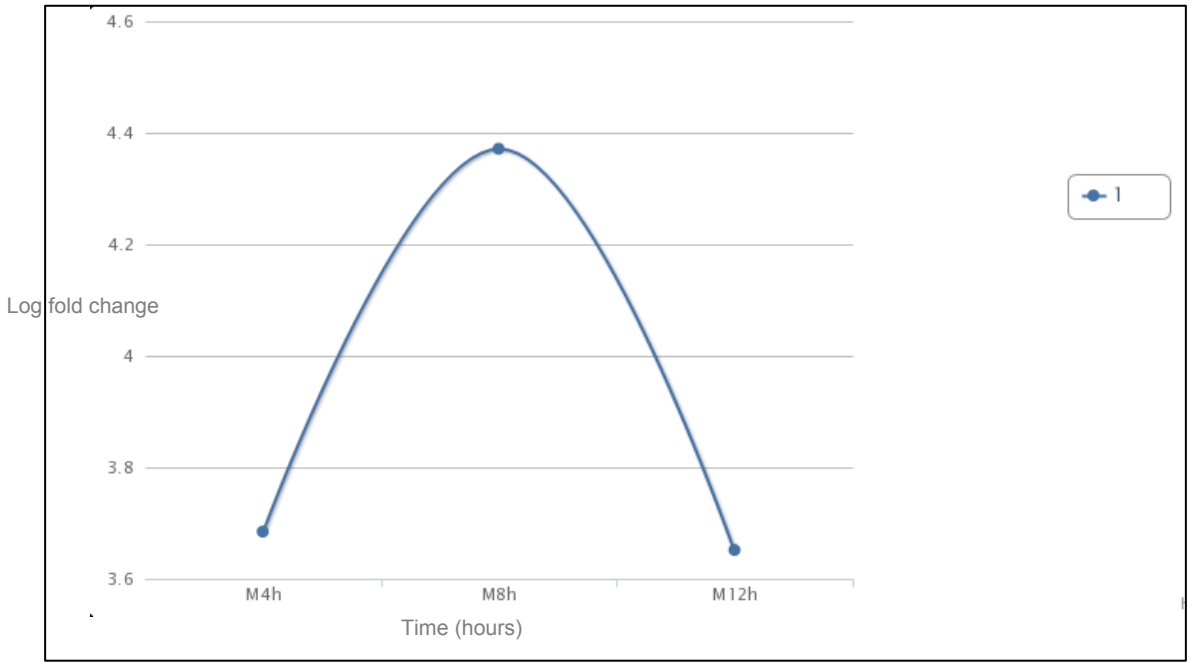


Figure 79 Macrophage data showing the fold change in macrophage (compared to control) at different time point in STM0018

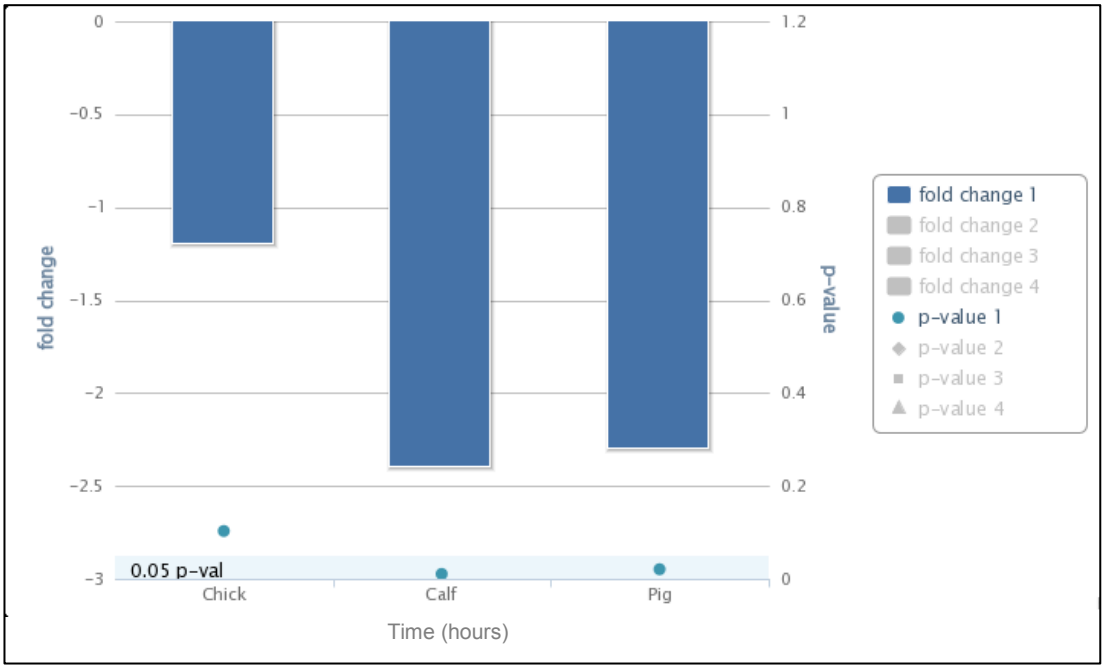


Figure 80 TraDIS data for mutation in STM0018 showing the significant negative selection in calves and pigs. The legend shows some data points faded out, this highlights the fact that the user can select which data points they want to see, in this diagram only the significant values are shown.

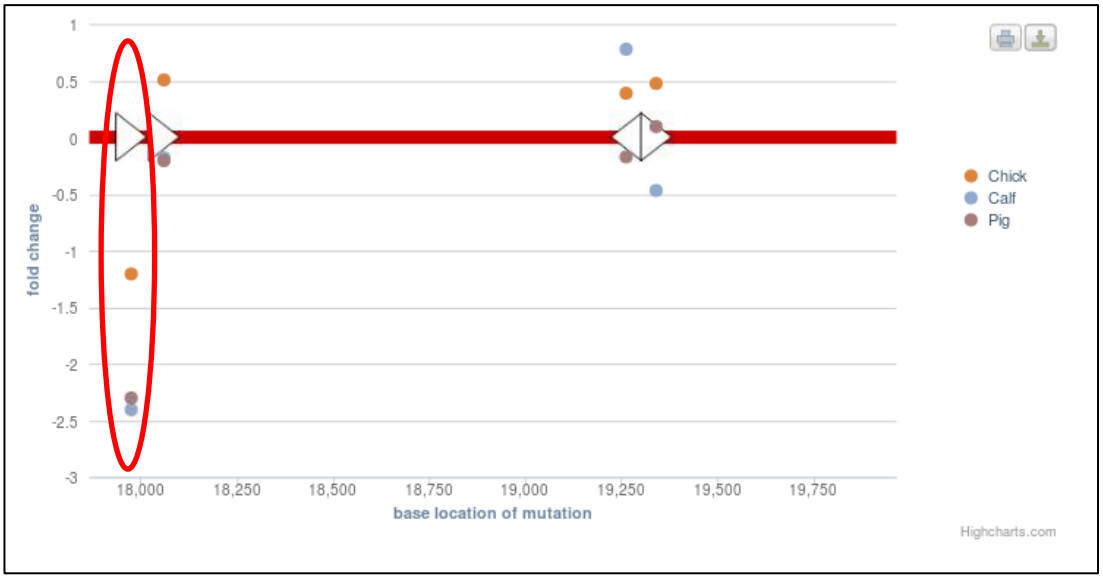


Figure 81 TraDIS context widget showing the mutation locations in STM0018. The triangles show the point of mutation and direction. The only significantly negatively selected mutation is circled in red.

Another example from the Caecal significant gene list is STM3764 (Figure 82) which codes for the MgtC protein, part of the $Mg2^{+}$ transporter operon. This gene is also up regulated in the Macrophage data and negatively selected across all hosts in the TraDIS data (Figure 83 and Figure 84). Previous studies have shown up regulation in human and mouse macrophage (Typhi and Typhimurium respectively), with the gene being required for growth in magnesium poor environments [237].The paper which supports the Caecal data, recognises that MgtC is up regulated in caeca but they connect this to redundant expression of the gene due to its location in that *Salmonella* pathogenicity island SPI-3. [94]. However, in a study to investigate the regulation of mgtC in serovar Typhi Retamal *et al.* found that the gene mutants showed impeded growth in human epithelial cells [237]. Further to this, the universal significant attenuation across all hosts supports the inference that the up regulation in caeca is associated with pathogenicity not only in macrophages but intestinal infection too.

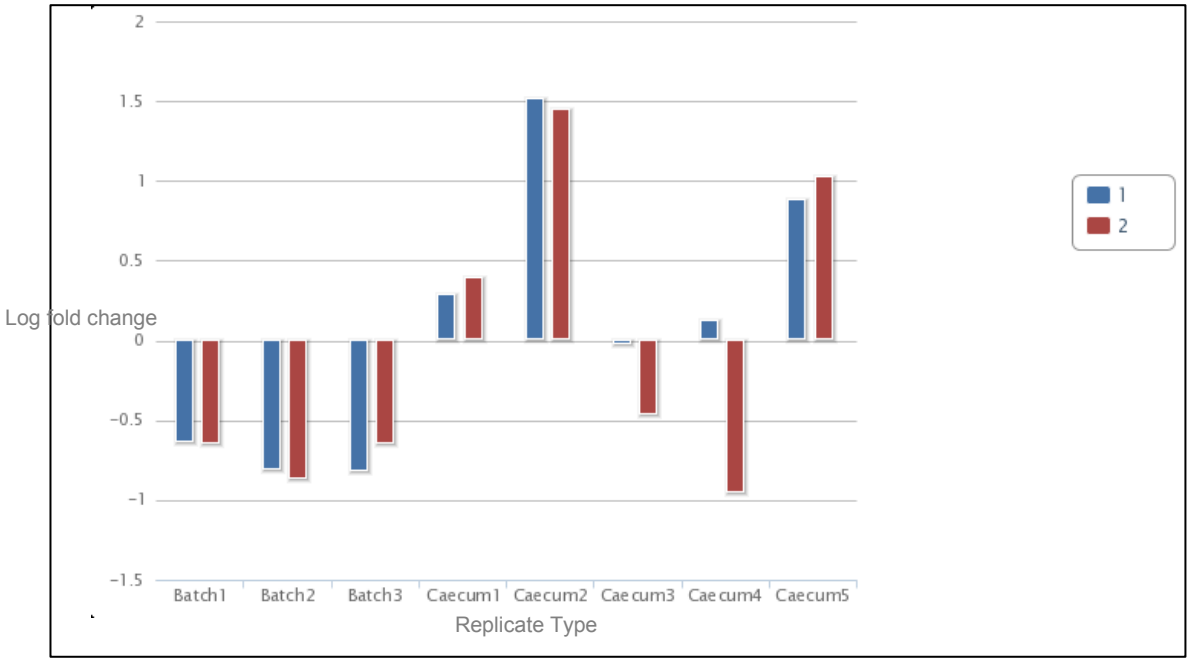


Figure 82 Caecal data for STM3764 there is up regulation in caeca when compared to LB Broth

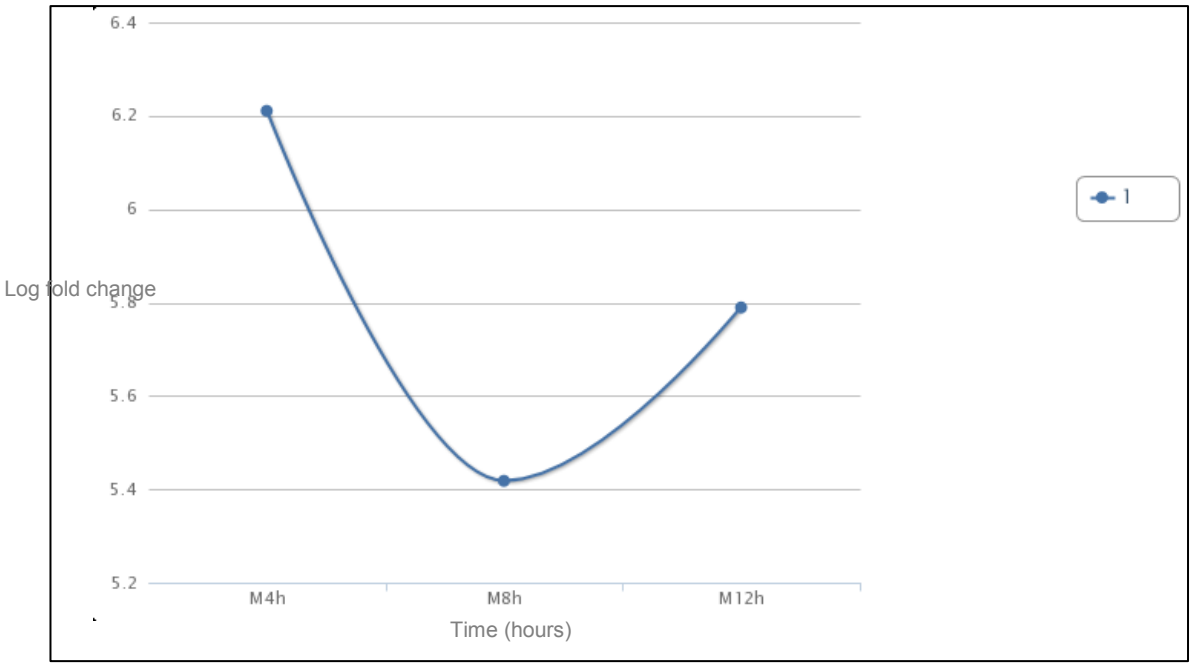


Figure 83 Macrophage Data for STM3764 showing significant up regulation of this gene in mouse macrophage when compared to the control

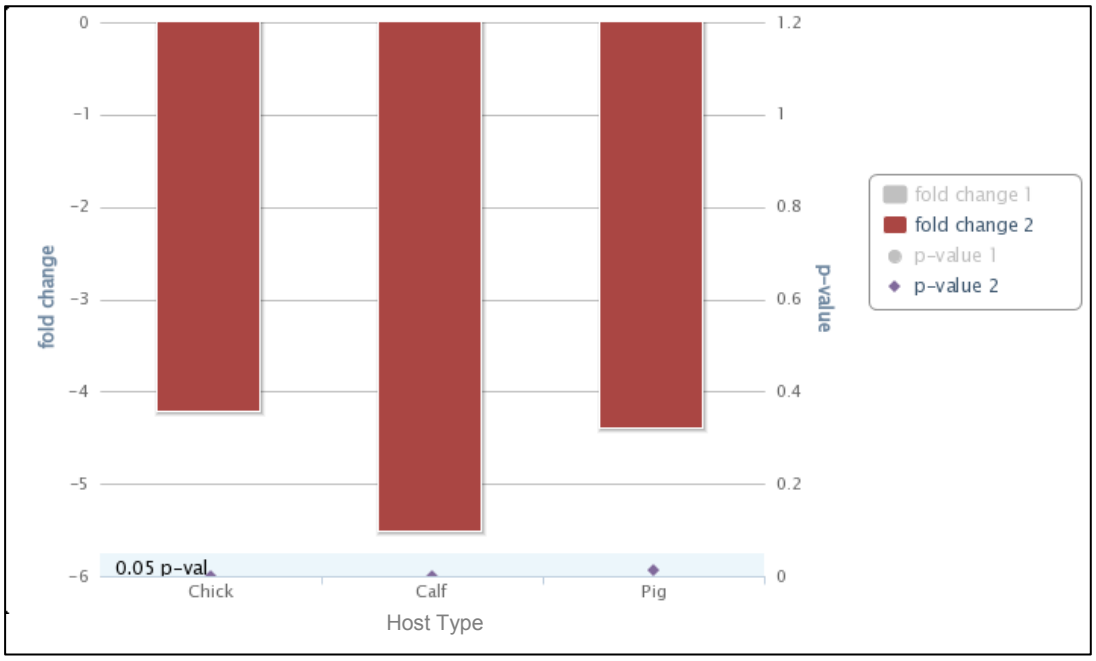


Figure 84 TraDIS data for the mutation in STM3764, showing significant negative selection across all hosts.

4.4 Discussion

GeneBook is a web-based bacterial genome browser that displays data from remote resources into a single instance. Users are able to customise their view making it relevant to their interests and personal datasets. The data available for not only *Salmonella* but many pathogens is disparate and there is a need for integration. Current enterobacterial resources hold their data locally. This is both memory intensive and the data is liable to become superseded. The PATRIC system [122] does continually update its information but it is constrained to holding the data locally. The PATRIC user interface does offer some personalisation with workspaces but there is still no option for integrating private data. At the time of starting this chapter (2009) none of the current resources had the functionality to upload and process raw sequences, now Xbase and PATRIC both offer an annotation service, although unlike GeneBook neither allow the integration of private genomes into their databases.

The fact that GeneBook draws information in live from any given data source means that it is as up-to-date as the primary source. This confers an advantage compared to other resources which rely on regular updates, sometimes from secondary data sources. This means that the quality of annotation for a feature can be assessed. Annotations that are out of date, misannotated, uninformative or unclear can be resolved allowing users to not only get a more holistic view of the feature of interest but also make informed decisions on the quality of an annotation. Overcoming these problems would normally rely on the use of multiple resources and running multiple tools, GeneBook does this automatically.

The in-house generic quantitative webservices can handle any quantitative data. The requirements are simple to adhere to, firstly tab delimited format and secondly, the use of GeneBook recognisable headers. Most text editors and spreadsheet software (such as Microsoft Excel) can save into tab delimited format and most scientific data has column headings which are easily edited. These webservices give users a quick method of visualising their own data in the context of other data. The alternative to this is to have multiple spreadsheets open.

In terms of handling experimental data GeneBook does not currently perform complex analyses, rather, it offers a robust foundation for hypothesis building. The advantage of seeing different experiments in the context of one another is that users can build hypotheses based on what the data indicates. Section 4.3 shows some of the hypotheses that can be made and what the next steps would be for a researcher in order to confirm their hypotheses.

This chapter has shown a method of integrating remote data sources allowing users to see many different types of data in one place. GeneBook integrates two microarray datasets, with the private TraDIS data along with public databases and data such as BLAST results, Domains, Orthologs and multiple alignments. These are all displayed from within a single browser window. This methodology is unique to GeneBook and impossible to achieve in any other single system. The methods such as a lightweight database and webservices are following the trends we see in web technology. There is general movement towards cloud computing and the paradigm of anyone accessing anything anywhere. GeneBook attempts to adhere to this model, relying on remote data sources rather than keeping the data locally.

4.5 Future work

The possibilities for future developments of GeneBook are endless due to the nature of it being customisable and extendible. In this section I have described some stages that I feel would be the next logical stage in its development.

4.5.1 Improving current widgets

The generic quantitative webservice can display p-values but I think that it could be enhanced by highlighting whether the results in the graph are significant. This could be as a note on the graph. It would allow a user to automatically see if the results from various experiments were significant. This could be achieved as a significance column or the p_value header could have an extra clause which is significance. That is the user would change the header to p_value_0.05 meaning anything lower than 0.05 would be highlighted as significant. This improvement would be especially useful for experiments that have multiple data points for each feature, users could see if the feature as a whole is significant.

One limitation of GeneBook in its current state is that some webservices rely on locus tags. This means that novel genomes which haven't been submitted to GenBank or Embl wouldn't get any results in these widgets as there is no publically available data for these genomes. I think that the way to improve this would be through orthology. All webservices would try the genome's locus tag and if this fails the ortholog's locus tag would be submitted. This could be achieved in two ways:

1) Extend the GeneBook database to include orthologs. This would be an extra entity that contains pairings of features that are calculated to be homologous. It would have a many to many relationship with the feature entity and use locus tag as the foreign key. The orthologs could be calculated by using reciprocal fasta against all the features available in GeneBook or by using the top hit from a BLAST search against nr.
2) Calculate orthologs on the fly, the sequence would be submitted to a alignment tool like BLAST, the top hit would be parsed for its ID which would then be used as the query.

Both options have their advantages and limitations. Option 1 would be faster as it is a simple query to the GeneBook database, however this method goes against the GeneBook model of a lightweight database with minimal information. Arguably a better ortholog might become available at a later date. Conversely, option 2 will be more up to date but it relies on a good internet connection and fast webservices, currently this could be a limitation but with the pace that internet speed is increasing there will become a point where the difference between running something locally and remotely will be negligible.

4.5.2 New widgets

With new tools and databases being developed constantly and a vast amount of resources already available the idea of developing a webservice to cover most areas of microbiology is well beyond the scope and timescale of this project. The paradigm of GeneBook is that if a user finds themselves repeatedly performing the same task they can include a new widget which performs this task automatically. This section

looks at some webservices that weren't developed but would be next given more time.

I described using e-utilities to access the CDD and return lists of domains that match the query feature. The next step would be to develop a webservice that can visualise domains in the context of the sequence. An example of it use would be in conjunction with pseudogene assessment to see whether domains are disrupted. I have already shown webservices in GeneBook that can visualise sequence data. Using pfamscan would give the location of domains which could then be visualised using the Bio::Graphics package in Perl.

Another widget I would like to develop would be a 'note box', allowing the user to make notes on the feature of interest that would be kept for future visits. This could be used to record discrepancies, perhaps links to literature of interest or some ideas. The obvious way of keeping the notes would be to add the records to a new entity in the database, this again, would not follow the GeneBook model but as the notes are limited to GeneBook perhaps this would be acceptable. However, a really exciting and novel method recording notes would be to develop a webservice that connects to note taking services like EverNote. In the case of Evernote this could be achieved via their cloud API [238], With this sort of webservice a user could make a note about a particular feature in the note taking software which could then be accessed via GeneBook when looking at the respective feature.

4.5.3 Next steps

If I had more time I would extend the search functionality of GeneBook. Rather than just looking at features users could look at a region of a genome either by giving two base locations or stating how far up and downstream of a feature they would like to look. This could rely on both locus tag/identifiers and sequence based queries. Regional searches would also entail some new webservices, perhaps looking for CRISPRs, tRNAs, small RNAs and other regulatory features that often aren't in the annotation.

The next major task in terms of searchability would be to allow genome wide searches. Allowing the user to submit a sequence as the query, which returns features/regions which are encompassed. Another genome wide search would be to query an external resource which returns a list of features that match the query. For example the user could submit a sequence this is queried against KEGG which returns a list of all the genes (within GeneBook) that are part of the same reaction in the same pathway. This list would then allow users to interrogate GeneBook based the results of their complex query. These genome wide searches could be extended to personal datasets, for example the scripts used for the enrichment analysis could be applied to certain datasets and bring back a list of enriched features.

4.5.4 Improving the system

There are several areas that would benefit from development that are not user-oriented, they are behind the scenes but affect operability.

A really interesting route with GeneBook would be to explore the possibility of not having a database at all. That means that all information would be obtained via webservices. For example the user would enter a protein ID of interest and genome ID. This query would shoot off to RefSeq and return the sequence for the given feature, it would also return the results from a protein database like UniProtKB. These results would then be sent off as queries against all the webservices that form the widgets in GeneBook. This would basically mean that GeneBook is a tool for integrating different web sources without holding any information at all. As far as the user is concerned the front end of GeneBook could be exactly the same, however behind this interface there would not be a database. One limitation of this method is that it would require a very fast internet connection and fast webservices as two waves of queries are fired off to remote data sources. However, this method would mean that genomes would never need to be uploaded to GeneBook, it could adapt to any publically available genome, or private genomes on an ftp site.

In conclusion, the state that GeneBook is in currently gives users the power to see multiple data types in the context of one another. There are many possibilities to

extend this resource and this section covers some diverse options. I believe that with the advances in internet speed the concept of developing GeneBook to hold no biological data is a feasible possibility and would serve to truly give users the most up to date information possible.

Chapter Five

Final conclusions

This chapter consolidates the previous chapters, giving an overview of the project and drawing conclusions from all of the pieces of work.

Bacterial genome sequencing has moved from sequencing one clonal bacterium to projects where thousands of bacteria are sequenced from the same population either a specific serovar or a metagenome to find bacterial variation or species distribution respectively. With this in mind the bottle neck in the process will be annotation and analysis. There are now annotation tools that can transfer annotation very quickly [151], but the problem still remains that relying on orthology means that the annotation is only as good as the genome it is transferred from. Also, if we are looking for differences and novelty in these genomes, these areas might be missed as by their very nature they are novel, there won't be any information to transfer across. The Shiga toxin in Escherichia coli O104:H4 (the strain that caused the outbreak in 2011) was missed during annotation transferral because it was in an HGT island. This was the crucial part of the genome that explained its high virulence. Only after careful review by multiple groups did they find the toxin [50]. In order to get a reasonable quality annotation there is still a requirement for manual correction/annotation. With a thousand genome scale study the feasibility of this is

low, compromising the quality of the overall annotation. In fact for most of these studies the genomes are never submitted to public repositories.

Further to the bottleneck created by big sequencing projects, there is also the deluge of quantitative post-genomic data. NGS technology has brought new technologies like RNA-seq and TraDIS in its wake. Although they have had teething problems, the experiments are at a point where the sequencing/experimental stage is negligible compared to the analysis.

Our sequencing effort has barely scratched the surface in terms of mass genome sequencing efforts. However the project does have merit. The four serovars, that we sequence in chapter 2, are of high biological interest, all isolated from infected hosts, important food-producing animals, all with known, well defined pathogenicity. These genomes can answer genotypic questions based on very specific phenotypes. With that in mind the fact that these are publicly available and more strains from the same serovars are being sequenced means that we have more scope for robust analyses of known traits.

S.enterica serves as a good model for pathogenicity studies. It is growing to be an important systemic human pathogen in developing countries and is associated with foodborne disease in the UK [6, 12]. It has a broad range of hosts and different serovars cause varying degrees of infection. The relationship between pathogenicity and host specificity is a complex one. The host pathogenicity model can be broadly summarised as servoars with a wide host range cause less severe infection and those which are restricted to one host cause severe systemic infection. Chapter 3 took the well-defined phenotypes from our serovars and used functional enrichment analysis of pseudogene formation in pathways and mutagenesis data to explore this model. This analysis forms a good foundation for building hypotheses around the *Salmonella* host specificity and pathogenicity model. For example, only the host generalist, *S. typhimurium* maintained had no pseudogenes/absences in Fructose metabolism, implying a link between gut colonisation and Fructose metabolism. Integrating this with the TraDIS data showed the Fructose pathway is essential for

Typhimurium to colonise all hosts. The next steps with this analysis is to test some of the hypotheses formed by performing knockouts.

The analysis of pathways is limited by the arbitrary method that we define them, the proof of concept for network analysis of the entire KEGG network showed that this type of analysis is hindered by file formatting and the fact that clustering of man-made networks tends to result in a skewed hairball structure, with a few nodes partaking in most of the interactions rather than the ideal network (for clustering purposes) that would have a similar frequency of interactions for each node. In order to perform an analysis that avoids the described bias in KEGG pathways, the analysis could be repeated using GO terms or a network of the TraDIS data could be made and the KEGG pathways could be mapped to this.

Assimilating different data at a genomic level is no mean feat, in terms of writing scripts, there are always exceptions to the defined rules which are missed and as previously discussed, orthology mapping isn't without its flaws. Even with its complexity, genome-scale analysis and integrating different datasets can lead to insights into Salmonella biology. This type of analysis is useful as it brings forth areas for hypothesis testing, but needs to be followed up by targeted experimental validation before any hard conclusions can be drawn.

The analysis in chapter 3 took a site wide approach to bacterial genomics. On the other hand chapter 4, the development of GeneBook took a feature level stance. The fact that a person is searching for a feature on a genome browser means that it is of some biological interest to them, whether they came across a gene in a paper or they have a feature that came up as significant during experimental validation. With the pace that data is being produced, it is not easy to keep up to date. GeneBook uses webservices to take the data directly from the primary source. Beyond being used as a general feature browser, user can upload their own data and see it in the context of public data and their other private datasets. Currently this type of activity would require multiple windows open compared to a simple point and click offered by GeneBook. Although there has been no formal feedback, GeneBook as has been

presented at several conferences and shown to colleagues. It has received good feedback with people recognising that it fills a unique niche. To take the GeneBook project forward beta-testing is required, with critical feedback being returned. The development of GeneBook so far doesn't offer a genome wide analysis, rather it serves as a means of viewing features of interest in the context of multiple, potentially remote, datasets. In terms of future work this is described in great detail in section 4.5, but one particular area of development that I would like to pursue is the direct enrichment analysis of data to provide lists of interest as a means of interrogating GeneBook, this would be achieved by modifying the scripts used in chapter 3 to be able to take generic quantitative data. In terms of priority areas of development for a beta-testing standard version I feel that automatic ortholog mapping will increase the functionality for genomes that have little data in the public domain, also increasing the functionality beyond basic gene feature to regions would increase the kinds of data available for GeneBook.

Another logical step forward would be towards complete use of webservices, that is no underlying database, thus truly following the dynamic up-to-date paradigm of GeneBook. This is a foreseeable step as we are on the verge of the super-fast internet era.

With the future of bacterial genomics moving towards thousands of genomes being sequenced we need to consider what happens to these genomes. Where are they stored, are they annotated? The manual aspect is no longer feasible. We are moving away from good quality annotations, the premise of gold standard genomes, outlined in section 2.4.3 could form a basis for improving quality. If this was integrated with webservices providing the most up to date information there would be less need for storing data in multiple places rather it would be accessed remotely when required reducing the risks of becoming obsolete and out of date. This project has shown the different aspects of bacterial genome analysis from sequencing and annotation of four serovars of biological interest through to analysis of the serovars and integration with other data type and finally visualisation of bacterial genomes in context of these different data sets. We have seen that the mechanisms behind pathogenicity and

specificity are complex, although there are no smoking guns this analysis has shown that it is possible to find patterns in pathways and patterns between mutagenesis in a host generalist and orthologs for the corresponding host using genome wide methods and a feature browser like GeneBook.

Appendix

Appendix A Files and Scripts

All of the files and scripts are located on the disk provided and can also be found at https://github.com/limeyloos/GeneBook, they are described in the README.txt file included with the files

Appendix B Full description of the submission process into GenBank

Submission process

The genomes were all annotated based on genomes available in GenBank and UniprotKB. Based on this our genomes were submitted to GenBank after they were annotated according to the method in 2.2.3. The sequences and their corresponding files were uploaded using the genome submission tool [153]

These annotations were not accepted by GenBank based on various annotation discrepancies (described in 2.3.1). The volume of these was so large that scripts were needed to process the discrepancies. Below are the steps used to make further submissions ultimately meeting GenBank's new annotation standards.

2nd Submission

Asn2disc, the NCBI programme for detecting annotation discrepancies (http://www.ncbi.nlm.nih.gov/GenBank/asndisc.html), was run against our primary submissions. Figure A1 shows an example of a discrepancy file summary.

All types of hypothetical proteins (including those labelled as ***'conserved hypothetical protein'*** and ***'putative uncharacterized protein'***) were extracted (Appendix A: get_hyp_prot.pl) and BLASTed against Swissprot and TREMBL.

The script ***parseBLAST2GenBank.pl*** was used to parse the BLAST results into the annotation（Appendix A: parseBLAST2GenBank.pl). Each hypothetical protein was checked for a hit (>85% across the length of the hit and the length query and >75% identity) to Swissprot, and failing that a hit to TREMBL, if there were no hits the annotation remained as a hypothetical protein (if the annotation was originally ***'conserved hypothetical protein'*** or ***'putative uncharacterized protein'*** these were converted to ***'hypothetical protein'***). The output consisted of both GenBank file format and tab-delimited. It is worth mentioning that this script adds a note to each new annotation that states

where the annotation was transferred from (e.g. "`similar to TREMBL D0ZIQ4_SALT1`").

```
Summary
DISC_SOURCE_QUALS_ASNDISC:taxname (all present, all unique)
DISC_FEATURE_COUNT:gene: 4518 present
DISC_FEATURE_COUNT:CDS: 4145 present
DISC_FEATURE_COUNT:tRNA: 75 present
DISC_FEATURE_COUNT:rRNA: 22 present
EXTRA_GENES:276 gene features are not associated with a CDS or RNA
feature.
DISC_COUNT_NUCLEOTIDES:1 nucleotide Bioseqs are present
GENE_PRODUCT_CONFLICT:47 coding regions have the same gene name as
another coding region but a different product.
DUPLICATE_GENE_LOCUS:71 genes have the same locus as another gene on the
same Bioseq.
EC_NUMBER_NOTE:1 features have EC numbers in notes or products.
OVERLAPPING_CDS:28 coding regions overlap another coding region with a
similar or identical name.
CONTAINED_CDS:42 coding regions are completely contained in another
coding region.
RNA_CDS_OVERLAP:9 coding regions overlap RNA features
SUSPECT_PRODUCT_NAMES:132 product_names contain 'suspect phrase or
characters'
       1 product names end with binding
       3 product names end with domain
       1 product names end with repeat
       1 product names contain 'Includes:'
       1 product names contain 'Salmonella'
       2 product names contain 'gene'
       1 product names contain 'utilisation'
       1 product names contain 'fold'
       5 product names contain 'three or more numbers together, not after
'UPF' or 'DUF' or 'IS' and not followed by the word 'family' and not
preceded by either 'cytochrome' or 'coenzyme''
       66 product names contain 'Brackets or parenthesis [] ()'
       1 product names contain 'containg'
       1 product names contain 'dependant'
       1 product names contain 'disulphide'
       2 product names contain 'golgi'
       1 product names contain 'haem'
       1 product names contain 'homeserine'
       1 product names contain 'hpothetical'
       1 product names contain 'hpothetical'
       1 product names contain 'puative'
       1 product names contain 'puative'
       1 product names contain 'putaive'
       1 product names contain 'putatve'
       1 product names contain 'signalling'
       2 product names contain 'sulpho'
       1 product names contain 'C-term'
       1 product names contain 'N-term'
       10 product names contain 'Two or more sets of brackets or
parentheseis'
       1 product names contain 'double space'
       2 product names contain '_'
       3 product names contain 'ending with period, comma, hyphen,
underscore, colon, or forward slash'
       15 product names contain 'may contain a plural'
       1 product names contain 'unbalanced brackets or parentheses'
N_RUNS:1 sequences have runs of 20 or more Ns
ADJACENT_PSEUDOGENES:3 pseudogenes match an adjacent pseudogene's text
```

Figure A1 Summary section from a discrep file produced by the software asn2disc. This shows the areas of genome annotation that need to be assessed for successful submission into GenBank. An example of a full discrepancy file is in Appendix A: full_discrepancy.txt

3rd Submission

Many of the annotations still had suspect names according to the discrepancy file but we believed that because they hit proteins in reference genomes with functional domains that these names would be acceptable. However, this was not the case, communication with the GenBank submission team returned this statement:

> *"It's an ongoing process to improve annotation on new submissions, rather than just duplicating poor annotation on old submissions. Therefore, we appreciate your help in modifying your submission to adhere to the current annotation goals"*

Figure A2 Excerpt from email communication with GenBank Sequence Submission staff explaining why transferred description can lead to poor annotation.

With the explanation from the GenBank submission staff in mind Asn2disc was run against this annotation to get discrepancies. The number of suspect proteins was so large that a script, ***get_bad_anno_prots.pl*** was written to extract the FASTA sequences for the suspect protein names (according to the discrepancy file) (Appendix A: get_bad_anno_prots.pl). These were then BLASTed and added to the annotation using ***parseBLAST2GenBank.pl*** (as described previously). Any suspect proteins which did not hit Swissprot or TREMBL within the previously stated thresholds were manually checked and corrected. This output was converted to .sqn as before and these sequences were then checked manually for the overlap discrepancies (RNA_CDS_OVERLAP, OVERLAPPING_CDS and CONTAINED_CDS). If the overlap discrepancies were the same as the reference genome or a hit to Swissprot/TREMBL then they were accepted. If the overlap discrepancies were not present in the reference genome and they had no domains (according to a pfam search [62]) they were removed from the annotation. This was then converted into .sqn (and the other appropriate files) again and submitted to GenBank.

4th Submission

All of the overlap discrepancies that were kept in the 3nd submission were present in reference genomes or UniprotKB. They were kept even if they were very short or did not have domain because of their high conservation to other genomes. Communication with the GenBank submission team (Figure) explained that overlapping proteins are rare in bacterial genomes and that conservation across genomes does not indicate a true CDS.

The script ***remove_bad_hyps.pl*** was written to take all the CDS with an overlap discrepancy and remove those which are less than 250bp and labelled as a hypothetical protein (Appendix A: remove_bad_hyps.pl). This reduced the list enough for manual annotation. The remaining overlapping proteins were checked for domains using pfam scan, if they were hypothetical and had no domains they were removed from the annotation. At this point any other discrepancies were checked manually, some discrepancies were kept in the annotation because upon closer inspection they weren't misannotations (explained in section 2.3.2.1).

"I have been discussing the Salmonella annotation with our Reference sequence staff. They are looking into correcting the reference genomes since these were annotated before we had all of the current checks in place. All of the proteins from these reference genomes were propogated into other databases, such as TREMBL. Therefore, matching something in TREMBL is not grounds for keeping it as that is an uncurated archive. We strongly recommend that you not include CDS features that are completely contained in other CDS features. There are some very large proteins in bacteria, but they are usually polyketide synthases and have domain hits that show this. A protein that has no domain hits (or is called "hypothetical protein") and overlaps a protein with a 'real' name is likely an artifact. Please review the overlapping CDS hits and only keep the ones that you believe are the most likely to be translated."

Figure A2 Excerpt from email communication with GenBank Sequence Submission staff explaining that the fact that 'protein' appears in TREMBL does not mean that it a true protein, especially if it overlaps other better annotated proteins.

Appendix C Pathways that have functional genes but no pseudogenes across any serovar

Description	Gene totals for each pathway				Expected pseudogene totals for each pathway			
	Gallinarum	Choleraesuis	Dublin	Typhimurium	Gallinarum	Choleraesuis	Dublin	Typhimurium
Nucleotide Metabolism	104	101	102	99	4	2	1	0
Purine metabolism	80	77	78	75	3	1	1	0
Replication and Repair	57	55	59	59	2	1	1	0
Pyrimidine metabolism	53	51	51	51	2	1	1	0
Folding, Sorting and Degradation	49	47	47	48	2	1	1	0
Citrate cycle (TCA cycle)	32	32	33	33	1	1	0	0
Aminoacyl-tRNA biosynthesis	28	27	28	28	1	0	0	0
Homopercous recombination	28	27	28	29	1	0	0	0
Lipopolysaccharide biosynthesis	27	27	27	27	1	0	0	0
Propanoate metabolism	25	24	24	26	1	0	0	0
Galactose metabolism	23	26	21	25	1	0	0	0
Phenylalanine, tyrosine and tryptophan biosynthesis	23	23	23	23	1	0	0	0
Mismatch repair	23	22	25	25	1	0	0	0
Glycerolipid metabolism	18	19	17	18	1	0	0	0
Ubiquinone and other terpenoid-quinone biosynthesis	18	18	18	18	1	0	0	0
Lysine biosynthesis	18	18	18	18	1	0	0	0
Protein export	18	17	16	18	1	0	0	0
DNA replication	18	16	19	19	1	0	0	0

Nicotinate and nicotinamide metabolism	16	16	17	16	1	0	0	0
Sulfur relay system	16	15	17	16	1	0	0	0
RNA degradation	15	15	14	14	1	0	0	0
Base excision repair	14	14	13	14	1	0	0	0
Folate biosynthesis	14	13	14	15	1	0	0	0
Histidine metabolism	13	14	13	13	1	0	0	0
One carbon pool by folate	13	11	14	12	1	0	0	0
Fatty acid biosynthesis	12	12	12	12	0	0	0	0
Terpenoid backbone biosynthesis	12	12	12	12	0	0	0	0
Biosynthesis of Other Secondary Metabolites	12	9	12	14	0	0	0	0
RNA polymerase	10	10	10	9	0	0	0	0
Transcription	10	10	10	9	0	0	0	0
Riboflavin metabolism	9	9	9	9	0	0	0	0
Nucleotide excision repair	8	8	8	8	0	0	0	0
Streptomycin biosynthesis	8	5	8	10	0	0	0	0
Lysine degradation	7	8	8	7	0	0	0	0
Biotin metabolism	6	7	7	6	0	0	0	0
Biosynthesis of unsaturated fatty acids	6	6	6	6	0	0	0	0
Biosynthesis of siderophore group nonribosomal peptides	6	6	6	6	0	0	0	0
Phenylalanine metabolism	5	6	6	6	0	0	0	0
D-Glutamine and D-glutamate metabolism	5	5	5	5	0	0	0	0

Polyketide sugar unit biosynthesis	5	2	5	5	0	0	0	0
Novobiocin biosynthesis	4	4	4	4	0	0	0	0
Phosphonate and phosphinate metabolism	4	4	4	4	0	0	0	0
D-Alanine metabolism	4	4	4	4	0	0	0	0
Taurine and hypotaurine metabolism	4	3	4	3	0	0	0	0
Aminobenzoate degradation	3	4	4	4	0	0	0	0
Dioxin degradation	3	3	3	3	0	0	0	0
Lipoic acid metabolism	3	3	3	2	0	0	0	0
Inositol phosphate metabolism	2	2	2	6	0	0	0	0
Sphingolipid metabolism	2	2	2	3	0	0	0	0
Arachidonic acid metabolism	2	2	2	2	0	0	0	0
Limonene and pinene degradation	2	2	2	2	0	0	0	0
Caprolactam degradation	2	2	2	2	0	0	0	0
Polycyclic aromatic hydrocarbon degradation	1	2	1	2	0	0	0	0

Appendix D Pathways that have pseudogenes across every serovar

pathway	description	Total gene counts for each pathway				Observed pseudogene counts for each pathway			
		Gallinarum	Choleraesuis	Dublin	Typhimurium	Gallinarum	Choleraesuis	Dublin	Typhimurium
2	**Metabolism**	705	690	695	712	22	11	10	3
01100	**Metabolic pathways**	650	637	639	652	21	9	10	2
8	**Carbohydrate Metabolism**	305	297	293	321	17	9	7	2
12	**Membrane Transport**	265	267	272	277	15	5	7	1
01110	**Biosynthesis of secondary metabolites**	275	268	270	272	7	1	1	2
01120	**Microbial metabolism in diverse environments**	196	188	190	201	4	3	2	1
1	**Amino Acid Metabolism**	206	207	201	206	10	2	1	2
18	**Energy Metabolism**	139	129	133	136	4	3	1	1
02060	**Phosphotransferase system (PTS)**	47	50	46	57	4	1	1	1
13	**Xenobiotics Biodegradation and Metabolism**	34	32	27	37	2	1	1	1
00910	**Nitrogen metabolism**	45	43	43	45	2	1	1	1

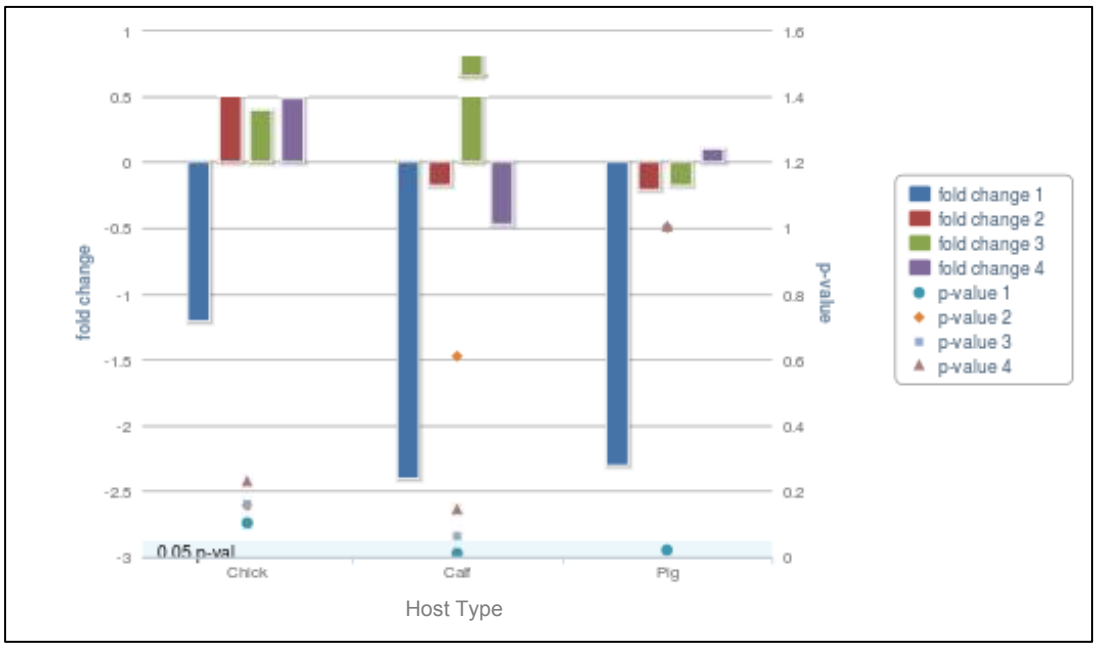
Appendix E Table of significantly attenuated mutations in pigs and show pseudogene/absence in Choleraesuis SCA50 orthology

LT2_locustag ortholog	Pseudogene in Choleraesuis?	Function	Pathway
STM0018	yes	Putative chitinase	
STM0032	ABSENT	arylsulfatase	
STM0033	ABSENT	5'-nucleotidase	
STM0035	YES	arylsulfatase	
STM0293	ABSENT	hypothetical protein	
STM0305	ABSENT	hypothetical protein	
STM0517	YES	glyoxylate carboligase (EC:4.1.1.47)	stm00630 Glyoxylate and dicarboxylate metabolism stm01100 Metabolic pathways
STM0723	ABSENT	ABC-type polysaccharide/polyol phosphate transport system ATPase component	Stm02010 ABC transporters
STM0810	YES	inner membrane protein	
STM0810	YES	inner membrane protein	
STM0810	YES	inner membrane protein	
STM0859	yes	LysR transcriptional regulator	
STM0859	yes	LysR transcriptional regulator	
STM0859	yes	LysR transcriptional regulator	
STM1018	ABSENT	hypothetical protein	
STM1092	ABSENT	hypothetical protein	
STM1092	ABSENT	hypothetical protein	

STM1094	ABSENT	pathogenicity island-encoded protein D	
STM1094	ABSENT	pathogenicity island-encoded protein D	
STM1332	ABSENT	O-antigen polymerase	
STM1555	ABSENT	transcriptional regulator	
STM1939	YES	glucose-6-phosphate dehydrogenase	
STM2189	YES	galactose/methyl galaxtoside transporter ATP-binding protein	stm02010 ABC transporters
STM2189	YES	galactose/methyl galaxtoside transporter ATP-binding protein	stm02010 ABC transporters
STM3083	yes	Putative mannitol dehydrogenase	
STM3254	ABSENT	fructose-1-phosphate kinase	stm00051 Fructose and mannose metabolism
STM3254	ABSENT	fructose-1-phosphate kinase	stm00051 Fructose and mannose metabolism
STM3255	ABSENT	phosphotransferase system fructose-specific component IIB (EC:2.7.1.69)	stm00051 Fructose and mannose metabolism stm01100 Metabolic pathways stm02060 Phosphotransferase system (PTS)
STM3255	ABSENT	phosphotransferase system fructose-specific component IIB (EC:2.7.1.69)	stm00051 Fructose and mannose metabolism stm01100 Metabolic pathways stm02060 Phosphotransferase system (PTS)
STM3638	yes	long polar fimbrial outer membrane usher protein	
STM4204	yes	inner membrane protein	
STM4204	yes	inner membrane protein	
STM4204	yes	inner membrane protein	
STM4204	yes	inner membrane protein	
STM4213	yes	Phage tail sheath protein	
STM4413	yes	metallo-dependent hydrolase	
STM4413	yes	metallo-dependent hydrolase	

STM4413	yes	metallo-dependent hydrolase	
STM474_0303	ABSENT	Rhs1 protein	

Appendix F Full TraDIS graph for STM0018



Appendix G The domains associated with STM0018 and their descriptions

cd06548 GH18_chitinase

The GH18 (glycosyl hydrolases, family 18) type II chitinases hydrolyze chitin, an abundant polymer of N-acetylglucosamine and have been identified in bacteria, fungi, insects, plants, viruses, and protozoan parasites. The structure of this domain is an eight-stranded alpha/beta barrel with a pronounced active-site cleft at the C-terminal end of the beta-barrel.

cl00046 ChtBD3

This group contains proteins related to the cellulose-binding domain of Erwinia chrysanthemi endoglucanase Z (EGZ) and Serratia marcescens chitinase B (ChiB). Gram negative plant parasite Erwinia chrysanthemi produces a variety of depolymerizing enzymes to metabolize pectin and cellulose on the host plant. Cellulase EGZ has a modular structure, with an N-terminal catalytic domain linked to a C-terminal cellulose-binding domain (CBD). CBD mediates the secretion activity of EGZ. Chitinases allow certain bacteria to utilize chitin as a energy source. Typically, non-plant chitinases are of the glycosidase family 18. Bacillus circulans Glycosidase ChiA1 hydrolyzes chitin and is comprised of several domains: the C-terminal chitin binding domain, an N-terminal catalytic domain, and 2 fibronectin type III-like domains. Bacillus circulans WL-12 ChiA1 facilitates invasion of fungal cell walls. The ChiA1 chitin binding domain is required for the specific recognition of insoluble chitin. although topologically and structurally related, ChiA1 lacks the characteristic aromatic residues of Erwinia chrysanthemi endoglucanase Z (CBD(EGZ)). Streptomyces griseus Chitinase C is a family 19 chitinase, and consists of a N-terminal chitin binding domain and a C-terminal chitin-catalytic domain that effects degradation. ChiC contains the characteristic chitin-binding aromatic residues. Chitinases function in invertebrates in the degradation of old exoskeletons, in fungi to utilize chitin in cell walls, and in bacteria which use chitin as an energy source.

References

1. Richardson, E.J., et al., *Genome sequences of Salmonella enterica serovar typhimurium, Choleraesuis, Dublin, and Gallinarum strains of well- defined virulence in food-producing animals.* J Bacteriol, 2011. **193**(12): p. 3162-3.
2. Richardson, E.J. and M. Watson, *The automatic annotation of bacterial genomes.* Brief Bioinform, 2012.
3. *DEFRA - Zoonoses report 2009*. Available from: http://archive.defra.gov.uk/foodfarm/farmanimal/diseases/atoz/zoonoses/documents/reports/zoonoses2009.pdf.
4. Wheeler, J.G., et al., *Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance. The Infectious Intestinal Disease Study Executive.* BMJ, 1999. **318**(7190): p. 1046-50.
5. Crump, J., S. Luby, and E. Mintz, *The global burden of typhoid fever.* Bulletin of the World Health Organisation, 2004. **82**: p. 346-353.
6. *Veterinary Laboratories Agency - Salmonella in Livestock Production in GB: 2011 Report*. 2011; Available from: http://vla.defra.gov.uk/reports/rep_salm_rep11.htm.
7. Deng W, L.S., Plunkett G(3rd), Mayhew GF, Rose DJ, Burland V, Kodoyianni V, Schwartz DC, Blattner FR, *Comparative genomics of Salmonella enterica serovar Typhi strains Ty2 and CT18.* Journal of Bacteriology, 2003. **185**: p. 2330-2337.
8. Uzzau, S., et al., *Host adapted serotypes of Salmonella enterica.* Epidemiology and Infection, 2000. **125**: p. 229-255.
9. Porwollik, S., et al., *Characterization of Salmonella enterica subspecies I genovars by use of microarrays.* Journal of Bacteriology, 2004. **186**: p. 5883-5898.
10. Haraga, A., M.B. Ohlson, and S.I. Miller, *Salmonellae interplay with host cells.* Nature Reviews Microbiology, 2008. **6**(1): p. 53-66.
11. Turner, A., et al., *Identification of Salmonella Typhimurium genes required for colonization of the chicken alimentary tract and for virulence in newly hatched chicks.* Infection Immunity, 1998. **66**: p. 2099-2106.
12. Stevens, M.P., T.J. Humphrey, and D.J. Maskell, *Molecular insights into farm animal and zoonotic Salmonella infections.*

Philos Trans R Soc Lond B Biol Sci, 2009. **364**(1530): p. 2709-23.
13. *Veterinary Laboratories Agency - Salmonella in Livestock Production in GB: 2009 Report. 2009 (chapter 1)*. 2009; Available from: http://vla.defra.gov.uk/reports/docs/rep_salm09_chp1.pdf.
14. Galan, J.E., *Molecular genetic bases of Salmonella entry into host cells.* Molecular Microbiology, 1996. **20**(2): p. 263-271.
15. Thompson, A., et al., *Salmonella transcriptomics: relating regulons, stimulons and regulatory networks to the process of infection.* Current Opinion in Microbiology, 2006. **9**: p. 109-116.
16. Wallis, T. and P. Barrow, *Salmonella epidemiology and pathogenesis in food-producing animals*, in *Escherichia coli and Salmonella: Cellular and Molecular Biology [online].* D. G, Editor. 2005, Washington DC: ASM Press.
17. Lawley, T., et al., *Genome-wide screen for Salmonella genes required for long-term systemic infection of the mouse.* PLoS Pathogens, 2006. **2**: p. e11.
18. Kothapalli, S., et al., *Diversity of genome structure in Salmonella enterica serovar typhi populations.* Journal of Bacteriology, 2005. **187**(8): p. 2638-2650.
19. Chan, K., et al., *Genomic comparison of Salmonella enterica serovars and Salmonella bongori by use of an S. enterica serovar typhimurium DNA microarray.* Journal of Bacteriology, 2003. **185**: p. 553-63.
20. Jacobsen, A., et al., *The Salmonella enterica pan-genome.* Microb Ecol, 2011. **62**(3): p. 487-504.
21. Hsiao, W., et al., *IslandPath: aiding detection of genomic islands in prokaryotes.* Bioinformatics, 2003. **19**(3): p. 418-420.
22. Holt KE, et al., *High throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi.* Nature Genetics, 2008. **40**: p. 987-993.
23. Rabsch, W., et al., *Salmonella enterica serotype Typhimurium and its host-adapted variants.* Infection Immunity, 2002. **70**: p. 2249-2255.
24. McClelland, M., et al., *Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid.* Nature Genetics, 2004. **36**(12): p. 1268-1274.

25. Lawrence, J.G., *Horizontal and vertical gene transfer: The life history of pathogens.* Contributions to Microbiology, 2005: p. 255-271.
26. Thomson, N., et al., *Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways.* Genome Research, 2008. **18**: p. 1624-1637.
27. Porwollik, S., et al., *Differences in gene content between Salmonella enterica serovar Enteritidis isolates and comparison to closely related serovars Gallinarum and Dublin.* Journal of Bacteriology, 2005. **187**: p. 6545-6555.
28. Nagy, G., et al., *Down-regulation of key virulence factors makes the Salmonella enterica serovar Typhimurium rfaH mutant a promising live-attenuated vaccine candidate.* Infection Immunity, 2006. **74**: p. 5914-5925.
29. Wilson, R.P., et al., *The Vi-capsule prevents Toll-like receptor 4 recognition of Salmonella.* Cellular Microbiology, 2008. **10**(4): p. 876-890.
30. Porwollik, S. and M. McClelland, *Lateral gene transfer in Salmonella.* Microbes and Infection, 2003. **5**(11): p. 977-989.
31. Chiu, C., et al., *The genome sequence of Salmonella enterica serovar Choleraesuis, a highly invasive and resistant zoonotic pathogen.* Nucleic Acids Research, 2005. **33**: p. 1690-1698.
32. Navarre, W.W., et al., *Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella.* Science, 2006. **313**(5784): p. 236-238.
33. Field, D., G. Wilson, and C. van der Gast, *How do we compare hundreds of bacterial genomes?* Current Opinion in Microbiology, 2006. **9**(5): p. 499-504.
34. Kyrpides, N. *GOLD Database Statistics*. 2008 [cited 2009 04/05]; Available from: http://genomesonline.org/gold_statistics.htm.
35. Galperin, M.Y. and G.R. Cochrane, *Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009.* Nucleic Acids Research, 2009. **37**(suppl_1): p. D1-4.
36. *Database: The Journal of Biological Databases and Curation*. 2009 [cited 2009 September]; Available from: http://database.oxfordjournals.org/.

37. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
38. Glenn, T.C., *Field guide to next-generation DNA sequencers.* Mol Ecol Resour, 2011. **11**(5): p. 759-69.
39. Loman, N.J., et al., *High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity.* Nat Rev Microbiol, 2012. **10**(9): p. 599-606.
40. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
41. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Hum Mol Genet, 2010. **19**(R2): p. R227-40.
42. Loman, N.J., et al., *Performance comparison of benchtop high-throughput sequencing platforms.* Nat Biotechnol, 2012. **30**(5): p. 434-9.
43. Gupta, P.K., *Single-molecule DNA sequencing technologies for future genomics research.* Trends Biotechnol, 2008. **26**(11): p. 602-11.
44. Maitra, R.D., J. Kim, and W.B. Dunbar, *Recent advances in nanopore sequencing.* Electrophoresis, 2012. **33**(23): p. 3418-28.
45. Eisenstein, M., *Oxford Nanopore announcement sets sequencing sector abuzz.* Nat Biotechnol, 2012. **30**(4): p. 295-6.
46. Ranieri, M., et al., *Comparison of Typing Methods with a New Procedure Based on Sequence Characterization for Salmonella Serovar Prediction.* Journal of Clinical Microbiology, 2013. **51**(6): p. 1786-1797.
47. Octavia, S. and R. Lan, *Single-nucleotide-polymorphism typing and genetic relationships of Salmonella enterica serovar typhi isolates.* Journal of Clinical Microbiology, 2007. **45**(11): p. 3795-3801.
48. Allard, M., et al., *High resolution clustering of Salmonella enterica serovar Montevideo strains using a next-generation sequencing approach.* Bmc Genomics, 2012. **13**.
49. Loman, N., et al., *A Culture-Independent Sequence-Based Metagenomics Approach to the Investigation of an Outbreak of Shiga-Toxigenic Escherichia coli O104:H4.* Jama-Journal of the American Medical Association, 2013. **309**(14): p. 1502-1510.

50. Rohde, H., et al., *Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4.* N Engl J Med, 2011. **365**(8): p. 718-24.
51. Hayward, M., V. Jansen, and M. Woodward, *Comparative genomics of Salmonella enterica serovars Derby and Mbandaka, two prevalent serovars associated with different livestock species in the UK.* Bmc Genomics, 2013. **14**.
52. Gordienko, E., M. Kazanov, and M. Gelfand, *Evolution of Pan-Genomes of Escherichia coli, Shigella spp., and Salmonella enterica.* Journal of Bacteriology, 2013. **195**(12): p. 2786-2792.
53. Magoc, T., et al., *GAGE-B: an evaluation of genome assemblers for bacterial organisms.* Bioinformatics, 2013. **29**(14): p. 1718-1725.
54. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
55. Milne, I., et al., *Tablet--next generation sequence assembly visualization.* Bioinformatics, 2010. **26**(3): p. 401-2.
56. Thorvaldsdóttir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2012.
57. MacLean, D., J.D. Jones, and D.J. Studholme, *Application of 'next-generation' sequencing technologies to microbial genetics.* Nat Rev Microbiol, 2009. **7**(4): p. 287-96.
58. Stothard, P. and D.S. Wishart, *Automated bacterial genome analysis and annotation.* Curr Opin Microbiol, 2006. **9**(5): p. 505-10.
59. Aggarwal, G. and R. Ramaswamy, *Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER.* Journal of Biosciences, 2002. **27**(1): p. 7-14.
60. Delcher, A.L., et al., *Improved microbial gene identification with GLIMMER.* Nucleic Acids Res, 1999. **27**(23): p. 4636-41.
61. Frishman, D., et al., *Combining diverse evidence for gene recognition in completely sequenced bacterial genomes.* Nucleic Acids Research, 1998. **26**(12): p. 2941-2947.
62. Finn, R.D., et al., *The Pfam protein families database.* Nucl. Acids Res., 2008. **36**(suppl_1): p. D281-288.
63. Pruitt, K.D. and D.R. Maglott, *RefSeq and LocusLink: NCBI gene-centered resources.* Nucleic Acids Research, 2001. **29**(1): p. 137-140.

64. Kanehisa, M. and S. Goto, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Research, 2000. **28**: p. 27-30.
65. Kaufmann, M., *The role of the COG database in comparative and functional genomics.* Current Bioinformatics, 2006. **1**(3): p. 291-300.
66. Aziz, R.K., et al., *The RAST Server: rapid annotations using subsystems technology.* BMC Genomics, 2008. **9**: p. 75.
67. Van Domselaar, G.H., et al., *BASys: a web server for automated bacterial genome annotation.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W455-9.
68. Lee, D., et al., *WeGAS: a web-based microbial genome annotation system.* Biosci Biotechnol Biochem, 2009. **73**(1): p. 213-6.
69. Vallenet, D., et al., *MaGe: a microbial genome annotation system supported by synteny results.* Nucleic Acids Res, 2006. **34**(1): p. 53-65.
70. Kumar, K., et al., *AGeS: a software system for microbial genome sequence annotation.* PLoS One, 2011. **6**(3): p. e17469.
71. Stewart, A.C., B. Osborne, and T.D. Read, *DIYA: a bacterial annotation pipeline for any genomics lab.* Bioinformatics, 2009. **25**(7): p. 962-3.
72. Yu, C., et al., *The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation.* BMC Bioinformatics, 2008. **9**: p. 52.
73. Cruveiller, S., et al., *MICheck: a web tool for fast checking of syntactic annotations of bacterial genomes.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W471-9.
74. Do, J.H. and D.K. Choi, *Computational approaches to gene prediction.* J Microbiol, 2006. **44**(2): p. 137-44.
75. Frishman, D., et al., *Combining diverse evidence for gene recognition in completely sequenced bacterial genomes.* Nucleic Acids Res, 1998. **26**(12): p. 2941-7.
76. Badger, J.H. and G.J. Olsen, *CRITICA: coding region identification tool invoking comparative analysis.* Mol Biol Evol, 1999. **16**(4): p. 512-24.
77. *Ongoing and future developments at the Universal Protein Resource.* Nucleic Acids Res, 2011. **39**(Database issue): p. D214-9.
78. Pearson, W.R., *Rapid and sensitive sequence comparison with FASTP and FASTA.* Methods Enzymol, 1990. **183**: p. 63-98.

79. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
80. Lowe, T.M. and S.R. Eddy, *tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.* Nucleic Acids Res, 1997. **25**(5): p. 955-64.
81. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
82. *The Bacterial Genome Submission Guide*. Available from: http://www.ncbi.nlm.nih.gov/genbank/genomesubmit.html.
83. *Genome Project Submission Account guidelines* Available from: http://www.ebi.ac.uk/embl/Submission/genomes.html.
84. Kummerfeld, S.K. and S.A. Teichmann, *Relative rates of gene fusion and fission in multi-domain proteins.* Trends Genet, 2005. **21**(1): p. 25-30.
85. Hollich, V. and E.L. Sonnhammer, *PfamAlyzer: domain-centric homology search.* Bioinformatics, 2007. **23**(24): p. 3382-3.
86. Koonin, E.V., *Orthologs, paralogs, and evolutionary genomics.* Annu Rev Genet, 2005. **39**: p. 309-38.
87. Kristensen, D.M., et al., *Computational methods for Gene Orthology inference.* Brief Bioinform, 2011. **12**(5): p. 379-91.
88. Overbeek, R., et al., *The use of gene clusters to infer functional coupling.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2896-901.
89. Forslund, K., I. Pekkari, and E.L. Sonnhammer, *Domain architecture conservation in orthologs.* BMC Bioinformatics, 2011. **12**: p. 326.
90. Sleator, R.D., C. Shortall, and C. Hill, *Metagenomics.* Letters in Applied Microbiology, 2008. **47**(5): p. 361-366.
91. Kurokawa, K., et al., *Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes.* DNA Research, 2007. **14**(4): p. 169-181.
92. Trevino, V., F. Falciani, and H. Barrera-Saldana, *DNA microarrays: a powerful genomic tool for biomedical and clinical research.* Molecular Medicine, 2007. **13**(9-10): p. 527-541.
93. Stekel, D., et al., *Analysis of host response to bacterial infection using error model based gene expression microarray experiments (vol 33, pg e53, 2005).* Nucleic Acids Research, 2005. **33**(7): p. 2352-2353.
94. Harvey, P.C., et al., *Salmonella enterica serovar typhimurium colonizing the lumen of the chicken intestine grows slowly and*

upregulates a unique set of virulence and metabolism genes. Infect Immun, 2011. **79**(10): p. 4105-21.
95. Wright, J.A., et al., *Multiple redundant stress resistance mechanisms are induced in Salmonella enterica serovar Typhimurium in response to alteration of the intracellular environment via TLR4 signalling.* Microbiology, 2009. **155**(Pt 9): p. 2919-29.
96. Mangan, M.W., et al., *The integration host factor (IHF) integrates stationary-phase and virulence gene expression in Salmonella enterica serovar Typhimurium.* Mol Microbiol, 2006. **59**(6): p. 1831-47.
97. Ball, C.A., et al., *Standards for Microarray data.* Science, 2002. **298**(5593): p. 539-539.
98. Barrett, T. and R. Edgar, *Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis.* Methods in Enzymology, 2006. **411**: p. 352-369.
99. Parkinson, H., et al., *ArrayExpress--a public database of microarray experiments and gene expression profiles.* Nucl. Acids Res., 2007. **35**(suppl_1): p. D747-750.
100. De Bruyne, V., F. Al-Mulla, and B. Pot, *Methods for microarray data analysis.* Methods Mol Biol, 2007. **382**: p. 373-91.
101. Quackenbush, J., *Microarray data normalization and transformation.* Nat Genet, 2002. **32 Suppl**: p. 496-501.
102. Pin, C., et al., *Comparison of different approaches for comparative genetic analysis using microarray hybridization.* Applied Microbiology and Biotechnology, 2006. **72**(4): p. 852-859.
103. Porwollik, S., R.M.-Y. Wong, and M. McClelland, *Evolutionary genomics of Salmonella: Gene acquisitions revealed by microarray analysis.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(13): p. 8956-8961.
104. Scaria, J., et al., *Microarray for molecular typing of Salmonella enterica serovars.* Molecular and Cellular Probes, 2008. **22**(4): p. 238-243.
105. Kim, H.-J., et al., *Microarray detection of food-borne pathogens using specific probes prepared by comparative genomics.* Biosensors and Bioelectronics, 2008. **24**(2): p. 238-246.
106. Hashimoto, K., et al., *KEGG as a glycome informatics resource.* Glycobiology, 2006. **16**(5): p. 63R-70R.

107. Chao, T. and N. Hansmeier, *The current state of microbial proteomics: Where we are and where we want to go.* Proteomics, 2012. **12**(4-5): p. 638-650.
108. Shi, L., et al., *Proteomic analysis of Salmonella enterica serovar Typhimurium isolated from RAW 264.7 macrophages - Identification of a novel protein that contributes to the replication of serovar Typhimurium inside macrophages.* Journal of Biological Chemistry, 2006. **281**(39): p. 29131-29140.
109. White, A., et al., *A Global Metabolic Shift Is Linked to Salmonella Multicellular Development.* Plos One, 2010. **5**(7).
110. Zhang, B. and R. Powers, *Analysis of bacterial biofilms using NMR-based metabolomics (vol 4, pg 1273, 2012).* Future Medicinal Chemistry, 2012. **4**(13): p. 1764-1764.
111. Maharjan, R. and T. Ferenci, *Global metabolite analysis: the influence of extraction methodology on metabolome profiles of Escherichia coli.* Analytical Biochemistry, 2003. **313**(1): p. 145-154.
112. Pan, Z. and D. Raftery, *Comparing and combining NMR spectroscopy and mass spectrometry in metabolomics.* Analytical and Bioanalytical Chemistry, 2007. **387**(2): p. 525-527.
113. Clayton DJ, et al., *Analysis of the role of 13 major fimbrial subunits in colonisation of the chicken intestines by Salmonella enterica serovar Enteritidis reveals a role for a novel locus.* BMC Microbiology, 2008. **8**(228).
114. Carnell, S., et al., *Role in virulence and protective efficacy in pigs of Salmonella enterica serovar Typhimurium secreted components identified by signature-tagged mutagenesis.* Microbiology, 2007. **153**: p. 1940-1952.
115. Morgan, E., et al., *Identification of host-specific colonization factors of Salmonella enterica serovar Typhimurium.* Molecular Microbiology, 2004. **54**: p. 994-1010.
116. Langridge, G.C., et al., *Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants.* Genome Res, 2009. **19**(12): p. 2308-16.
117. Chaudhuri, R., A. Khan, and M. Pallen, *coliBASE: an online database for Escherichia coli, Shigella and Salmonella comparative genomics.* Nucleic Acids Research, 2004. **32**: p. D296-D299.

118. Chaudhuri, R. and M. Pallen, *xBASE, a collection of online databases for bacterial comparative genomics.* Nucleic Acids Research, 2006. **34**: p. D335-D337.
119. Glasner, J., et al., *Enteropathogen Resource Integration Center (ERIC): bioinformatics support for research on biodefense-relevant enterobacteria.* Nucleic Acids Research, 2007. **36**: p. D519-D523.
120. Markowitz, V.M., et al., *IMG: the Integrated Microbial Genomes database and comparative analysis system.* Nucleic Acids Res, 2012. **40**(Database issue): p. D115-22.
121. Hertz-Fowler, C., et al., *GeneDB: a resource for prokaryotic and eukaryotic organisms.* Nucleic Acids Research, 2004. **32**: p. D339-D343.
122. Gillespie, J.J., et al., *PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species.* Infect Immun, 2011. **79**(11): p. 4286-98.
123. Sun, S., et al., *Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource.* Nucleic Acids Res, 2011. **39**(Database issue): p. D546-51.
124. *World Wide Web Consortium (W3C)*. Available from: http://www.w3.org/.
125. Neerincx, P.B.T. and J.A.M. Leunissen, *Evolution of web services in bioinformatics.* Briefings in Bioinformatics, 2005. **6**(2): p. 178-188.
126. Smedley, D., et al., *BioMart - biological queries made easy.* BMC Genomics, 2009. **10**(1): p. 22.
127. McWilliam, H., et al., *Web services at the European Bioinformatics Institute-2009.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W6-10.
128. *KEGG SOAP API*. Available from: http://www.genome.jp/kegg/soap/.
129. *Ensembl REST API – Ensembl Data from Any Language*. Available from: http://www.ensembl.info/blog/2012/09/27/ensembl-rest-api-ensembl-data-from-any-language/.
130. Rodriguez-Tome, P., *The BioCatalog.* Bioinformatics, 1998. **14**(5): p. 469-470.
131. Oinn, T., et al., *Taverna: a tool for the composition and enactment of bioinformatics workflows.* Bioinformatics, 2004. **20**(17): p. 3045-3054.

132. Goble, C.A., et al., *myExperiment: a repository and social network for the sharing of bioinformatics workflows.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W677-82.
133. Eriksson, S., et al., *Unravelling the biology of macrophage infection by gene expression profiling of intracellular Salmonella enterica.* Mol Microbiol, 2003. **47**(1): p. 103-18.
134. Jones, P.W., P. Collins, and M.M. Aitken, *Passive protection of calves against experimental infection with Salmonella typhimurium.* Vet Rec, 1988. **123**(21): p. 536-41.
135. Rankin, J.D. and R.J. Taylor, *The estimation of doses of Salmonella typhimurium suitable for the experimental production of disease in calves.* Vet Rec, 1966. **78**(21): p. 706-7.
136. Iqbal, M., et al., *Identification and functional characterization of chicken toll-like receptor 5 reveals a fundamental role in the biology of infection with Salmonella enterica serovar typhimurium.* Infect Immun, 2005. **73**(4): p. 2344-50.
137. Paulin, S.M., et al., *Net replication of Salmonella enterica serovars Typhimurium and Choleraesuis in porcine intestinal mucosa and nodes is associated with their differential virulence.* Infect Immun, 2007. **75**(8): p. 3950-60.
138. Bolton, A.J., et al., *Interaction of Salmonella choleraesuis, Salmonella dublin and Salmonella typhimurium with porcine and bovine terminal ileum in vivo.* Microbiology, 1999. **145 (Pt 9)**: p. 2431-41.
139. Paulin, S.M., et al., *Analysis of Salmonella enterica serotype-host specificity in calves: avirulence of S. enterica serotype gallinarum correlates with bacterial dissemination from mesenteric lymph nodes and persistence in vivo.* Infect Immun, 2002. **70**(12): p. 6788-97.
140. Bolton, A.J., et al., *Invasiveness of Salmonella serotypes Typhimurium, Choleraesuis and Dublin for rabbit terminal ileum in vitro.* J Med Microbiol, 1999. **48**(9): p. 801-10.
141. Wallis, T.S., et al., *The Salmonella dublin virulence plasmid mediates systemic but not enteric phases of salmonellosis in cattle.* Infect Immun, 1995. **63**(7): p. 2755-61.
142. SMITH, H.W., *Observations on experimental fowl typhoid.* J Comp Pathol, 1955. **65**(1): p. 37-54.
143. Bentley, D., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-59.

144. Li, R., et al., *De novo assembly of human genomes with massively parallel short read sequencing.* Genome Res, 2010. **20**(2): p. 265-72.
145. Kröger, C., et al., *The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium.* Proc Natl Acad Sci U S A, 2012. **109**(20): p. E1277-86.
146. Yu, H., et al., *Complete nucleotide sequence of pSCV50, the virulence plasmid of Salmonella enterica serovar Choleraesuis SC-B67.* Plasmid, 2006. **55**(2): p. 145-51.
147. Kurtz, S., et al., *Versatile and open software for comparing large genomes.* Genome Biol, 2004. **5**(2): p. R12.
148. Bonfield, J.K., K. Smith, and R. Staden, *A new DNA sequence assembly program.* Nucleic Acids Res, 1995. **23**(24): p. 4992-9.
149. Schattner, P., A.N. Brooks, and T.M. Lowe, *The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W686-9.
150. Suzek, B.E., et al., *A probabilistic method for identifying start codons in bacterial genomes.* Bioinformatics, 2001. **17**(12): p. 1123-30.
151. Otto, T.D., et al., *RATT: Rapid Annotation Transfer Tool.* Nucleic Acids Res, 2011. **39**(9): p. e57.
152. *NetBLASt*. Available from: http://www.biology.wustl.edu/gcg/netblast.html.
153. *NCBI - Genomes MacroSend*. Available from: http://www.ncbi.nlm.nih.gov/projects/GenomeSubmit/genome_submit.cgi.
154. Bergthorsson, U. and J.R. Roth, *Natural isolates of Salmonella enterica serovar Dublin carry a single nadA missense mutation.* J Bacteriol, 2005. **187**(1): p. 400-3.
155. Wang, X., et al., *Estimation of sequencing error rates in short reads.* Bmc Bioinformatics, 2012. **13**.
156. *Lucene – java based search engine*. Available from: http://lucene.apache.org/java/docs/index.html
157. *PHP class – 'did you mean?'*. Available from: http://www.phpclasses.org/package/4569-PHP-Get-spelling-correction-suggestions-from-Google.html
158. Rudd, K.E., *Linkage map of Escherichia coli K-12, edition 10: the physical map.* Microbiol Mol Biol Rev, 1998. **62**(3): p. 985-1019.

159. Gilks, W.R., et al., *Percolation of annotation errors through hierarchically structured protein sequence databases.* Math Biosci, 2005. **193**(2): p. 223-34.
160. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D211-22.
161. Bateman, A., P. Coggill, and R.D. Finn, *DUFs: families in search of function.* Acta Crystallogr Sect F Struct Biol Cryst Commun, 2010. **66**(Pt 10): p. 1148-52.
162. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
163. Saenz, H.L. and C. Dehio, *Signature-tagged mutagenesis: technical advances in a negative selection method for virulence gene identification.* Curr Opin Microbiol, 2005. **8**(5): p. 612-9.
164. *NCBI Complete Microbial Genomes*. Available from: www.ncbi.nlm.nih.gov/genomes/lproks.cgi.
165. Luo, C., G.Q. Hu, and H. Zhu, *Genome reannotation of Escherichia coli CFT073 with new insights into virulence.* BMC Genomics, 2009. **10**: p. 552.
166. Gundogdu, O., et al., *Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence.* BMC Genomics, 2007. **8**: p. 162.
167. Camus, J.C., et al., *Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv.* Microbiology, 2002. **148**(Pt 10): p. 2967-73.
168. Janssen, P., et al., *Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications.* EMBO Rep, 2005. **6**(5): p. 397-9.
169. Barrell, D., et al., *The GOA database in 2009--an integrated Gene Ontology Annotation resource.* Nucleic Acids Res, 2009. **37**(Database issue): p. D396-403.
170. *Webin EMBL-EBI annotation features and qualifiers*. Available from: http://www.ebi.ac.uk/ena/WebFeat/
171. Suzek, B.E., et al., *A probabilistic method for identifying start codons in bacterial genomes.* Bioinformatics, 2001. **17**(12): p. 1123-1130.
172. Ermolaeva, M.D., et al., *Prediction of transcription terminators in bacterial genomes.* Journal of Molecular Biology, 2000. **301**(1): p. 27-33.

173. Sigrist, C.J., et al., *PROSITE, a protein domain database for functional characterization and annotation.* Nucleic Acids Res, 2010. **38**(Database issue): p. D161-6.
174. Attwood, T.K., et al., *PRINTS and its automatic supplement, prePRINTS.* Nucleic Acids Res, 2003. **31**(1): p. 400-2.
175. Mulder, N. and R. Apweiler, *InterPro and InterProScan: tools for protein sequence classification and comparison.* Methods Mol Biol, 2007. **396**: p. 59-70.
176. Hacker, J., et al., *Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution.* Molecular Microbiology, 1997. **23**(6): p. 1089-1097.
177. Waack, S., et al., *Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models.* BMC Bioinformatics, 2006. **7**(1): p. 142.
178. Langille, M., W. Hsiao, and F. Brinkman, *Evaluation of genomic island predictors using a comparative genomics approach.* BMC Bioinformatics, 2008. **9**(1): p. 329.
179. Barrangou, R., et al., *CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes.* Science, 2007. **315**(5819): p. 1709-1712.
180. Kassai-Jáger, E., et al., *Distribution and evolution of short tandem repeats in closely related bacterial genomes.* Gene, 2008. **410**(1): p. 18-25.
181. Bland, C., et al., *CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats.* BMC Bioinformatics, 2007. **8**(209): p. 209.
182. Grissa, I., et al., *CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W52-7.
183. Sreenu, V.B., et al., *MICdb: database of prokaryotic microsatellites.* Nucleic Acids Research, 2003. **31**(1): p. 106-108.
184. Lu, Z., et al., *Predicting subcellular localization of proteins using machine-learned classifiers.* Bioinformatics, 2004. **20**(4): p. 547-556.
185. Hua, S. and Z. Sun, *Support vector machine approach for protein subcellular localization prediction.* Bioinformatics, 2001. **17**(8): p. 721-728.
186. Yu, C.S., C.J. Lin, and J.K. Hwang, *Predicting subcellular localization of proteins for Gram-negative bacteria by support*

vector machines based on n-peptide compositions. Protein Sci, 2004. **13**(5): p. 1402-6.
187. Wang, J., et al., *Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines.* BMC Bioinformatics, 2005. **6**(1): p. 174.
188. Gardy, J.L., et al., *PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis.* Bioinformatics, 2005. **21**(5): p. 617-623.
189. Gardy, J.L. and F.S.L. Brinkman, *Methods for predicting bacterial protein subcellular localization.* Nat Rev Micro, 2006. **4**(10): p. 741-751.
190. Watson, M., *ProGenExpress: visualization of quantitative data on prokaryotic genomes.* BMC Bioinformatics, 2005. **6**: p. 98.
191. Bhagat, J., et al., *BioCatalogue: a universal catalogue of web services for the life sciences.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W689-94.
192. Zhang, W., et al., *A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies.* PLoS One, 2011. **6**(3): p. e17915.
193. Vezzi, F., G. Narzisi, and B. Mishra, *Feature-by-feature--evaluating de novo sequence assembly.* PLoS One, 2012. **7**(2): p. e31002.
194. Wu, X. and M. Watson, *CORNA: testing gene lists for regulation by microRNAs.* Bioinformatics, 2009. **25**(6): p. 832-3.
195. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological)*, 1995. **57**(1): p. 289-300.
196. Theocharidis, A., et al., *Network visualization and analysis of gene expression data using BioLayout Express(3D).* Nature Protocols, 2009. **4**(10): p. 1535-1550.
197. Beste, D., et al., *The Genetic Requirements for Fast and Slow Growth in Mycobacteria.* Plos One, 2009. **4**(4).
198. Beste, D. and J. McFadden, *System-level strategies for studying the metabolism of Mycobacterium tuberculosis.* Molecular Biosystems, 2010. **6**(12): p. 2363-2372.

199. Beste, D. and J. McFadden, *Systems biology of the metabolism of Mycobacterium tuberculosis.* Biochemical Society Transactions, 2010. **38**: p. 1286-1289.
200. Sroka, J., et al., *Acorn: A grid computing system for constraint based modeling and visualization of the genome scale metabolic reaction networks via a web interface.* Bmc Bioinformatics, 2011. **12**.
201. Fricke, W., et al., *Comparative Genomics of 28 Salmonella enterica Isolates: Evidence for CRISPR-Mediated Adaptive Sublineage Evolution.* Journal of Bacteriology, 2011. **193**(14): p. 3556-3568.
202. Boyen, F., et al., *The fibronectin binding protein ShdA is not a prerequisite for long term faecal shedding of Salmonella typhimurium in pigs.* Vet Microbiol, 2006. **115**(1-3): p. 284-90.
203. Stincone, A., et al., *A systems biology approach sheds new light on Escherichia coli acid resistance.* Nucleic Acids Research, 2011. **39**(17): p. 7512-7528.
204. Becker, D., et al., *Robust Salmonella metabolism limits possibilities for new antimicrobials.* Nature, 2006. **440**(7082): p. 303-7.
205. Bumann, D., *System-level analysis of Salmonella metabolism during infection.* Curr Opin Microbiol, 2009. **12**(5): p. 559-67.
206. McMeechan, A., et al., *Glycogen production by different Salmonella enterica serotypes: contribution of functional glgC to virulence, intestinal colonization and environmental survival.* Microbiology, 2005. **151**(Pt 12): p. 3969-77.
207. Morgan, X.C., et al., *Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment.* Genome Biol, 2012. **13**(9): p. R79.
208. Chaudhuri, R., et al., *Comprehensive Assignment of Roles for Salmonella Typhimurium Genes in Intestinal Colonization of Food-Producing Animals.* Plos Genetics, 2013. **9**(4).
209. Hartemink, R., K.M. Van Laere, and F.M. Rombouts, *Growth of enterobacteria on fructo-oligosaccharides.* J Appl Microbiol, 1997. **83**(3): p. 367-74.
210. Monnier, V.M., *Bacterial enzymes that can deglycate glucose- and fructose-modified lysine.* Biochem J, 2005. **392**(Pt 2): p. e1-3.

211. Ten Bruggencate, S.J., et al., *Dietary fructo-oligosaccharides and inulin decrease resistance of rats to salmonella: protective role of calcium.* Gut, 2004. **53**(4): p. 530-5.
212. Pacheco, A.R., et al., *Fucose sensing regulates bacterial intestinal colonization.* Nature, 2012. **492**(7427): p. 113-7.
213. Li, J., et al., *Evolutionary origin and radiation of the avian-adapted non-motile salmonellae.* J Med Microbiol, 1993. **38**(2): p. 129-39.
214. Birney, E. *Ewan's Blog - 10 rules-of-thumb in genomics*. 2011; Available from: http://genomeinformatician.blogspot.co.uk/2011/07/10-rules-of-thumb-in-genomics.html.
215. *GUS - The Genomics Unified Schema*. [cited 2009 03/03]; Available from: http://www.gusdb.org/.
216. *Introduction to Chado*. 2008; Available from: http://gmod.org/wiki/Chado.
217. *BioSQL Wiki*. 2009 [cited 2009 25/02]; This is the main resource for BioSQL]. Available from: http://www.biosql.org.
218. *Drupal*. Available from: http://drupal.org/.
219. *Drupal - Homebox*. Available from: http://drupal.org/project/homebox.
220. *jQuery - The write less do more library*. Available from: http://jquery.com/.
221. Bairoch, A., *The ENZYME database in 2000.* Nucleic Acids Res, 2000. **28**(1): p. 304-5.
222. Consortium, U., *Reorganizing the protein space at the Universal Protein Resource (UniProt).* Nucleic Acids Res, 2012. **40**(Database issue): p. D71-5.
223. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets--update.* Nucleic Acids Res, 2012.
224. *Jalview - Applet*. Available from: http://www.jalview.org/examples/applets.html.
225. Waterhouse, A.M., et al., *Jalview Version 2--a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-91.
226. *Entrez Programming Utilities Help*. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25501/.
227. *Conserved Domains and Protein Classification*. Available from: http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml.

228. *Bio::Graphics - Generate GD images of Bio::Seq objects*. Available from: http://search.cpan.org/~lds/Bio-Graphics-2.32/lib/Bio/Graphics.pm.
229. Durinck, S., et al., *GenomeGraphs: integrated genomic data visualization with R.* BMC Bioinformatics, 2009. **10**: p. 2.
230. *Graphviz - Graph Visualization Software*. Available from: http://www.graphviz.org/.
231. Yin, T., D. Cook, and M. Lawrence, *ggbio: an R package for extending the grammar of graphics for genomic data.* Genome Biol, 2012. **13**(8): p. R77.
232. ***Highcharts** - Interactive JavaScript charts for your webpage*. Available from: http://www.highcharts.com/.
233. Dong, H., et al., *Roles of the spiA gene from Salmonella enteritidis in biofilm formation and virulence.* Microbiology, 2011. **157**(Pt 6): p. 1798-805.
234. Kisiela, D.I., et al., *Evolution of Salmonella enterica virulence via point mutations in the fimbrial adhesin.* PLoS Pathog, 2012. **8**(6): p. e1002733.
235. Frye, J., et al., *Identification of new flagellar genes of Salmonella enterica serovar Typhimurium.* J Bacteriol, 2006. **188**(6): p. 2233-43.
236. Larsen, T., et al., *Characterization of a novel Salmonella Typhimurium chitinase which hydrolyzes chitin, chitooligosaccharides and an N-acetyllactosamine conjugate.* Glycobiology, 2011. **21**(4): p. 426-36.
237. Retamal, P., M. Castillo-Ruiz, and G.C. Mora, *Characterization of MgtC, a virulence factor of Salmonella enterica Serovar Typhi.* PLoS One, 2009. **4**(5): p. e5551.
238. *Evernote Cloud API*. Available from: http://dev.evernote.com/documentation/cloud/.